Informatiser le lexique

Benoît Sagot

École Doctorale Informatique, Télécommunication et Électronique

Habilitation à Diriger les Recherches

Discipline : Informatique

Présentée et soutenue publiquement
par

Benoît Sagot

le 28 juin 2018

Informatiser le lexique

Modélisation, développement et exploitation de lexiques morphologiques, syntaxiques et sémantiques

Composition du jury :

Philippe Blache	CNRS (France), rapporteur
James P. Blevins	Université de Cambridge (Royaume-Uni), rapporteur
Ludovic Denoyer	Sorbonne Université (Paris, France)
Christiane D. Fellbaum	Université de Princeton (New-York, États-Unis), rapporteure
Anna Korhonen	Université de Cambridge (Royaume-Uni)
Laurent Romary	Inria (France), parrain
Gertjan Van Noord	Université de Groningue (Pays-Bas)

Table des matières

Introduction

Sommaire

Les recherches en traitement automatique des langues sont aussi anciennes que l'informatique elle-même et constituent dès les années 1950 une dimension-clef du domaine naissant de l'intelligence artificielle. Certaines idées et propositions concrètes, comme la machine à traduire de Troyanski (1935), remontent à la première moitié du vingtième siècle, voire aux siècles passés, mais c'est bien au début des années 1950 que le domaine émerge réellement. Les premières années ont été dédiées avant tout à la traduction automatique, au moyen de systèmes de règles et de dictionnaires électroniques. Pendant plusieurs décennies, les recherches en traitement automatique des langues ont eu pour objet le développement de modèles formels et computationnels des langues. Ces recherches concernaient majoritairement la langue anglaise, ce qui a lourdement influencé le domaine, et ce pas toujours positivement, tout en conduisant à des contributions cruciales en informatique théorique, notamment dans le domaine des grammaires formelles. Le rôle du lexique, central dans les premières expériences de traduction automatique, a connu un renouveau dans les années 1970 et 1980 avec l'émergence de systèmes formels dits lexicalisés pour la représentation des structures syntaxiques (Joshi *et al.*, 1975 ; Gross, 1975 ; Bresnan, 1982 ; Gazdar, 1985). Ces approches reconnaissaient dans le lexique le lieu adapté à l'encodage des faits linguistiques spécifiques aux mots individuels, conjointement à la grammaire, adaptée quant à elle à l'encodage des faits linguistiques génériques.

Dans ce contexte, les années 1990 ont vu émerger, en tout cas pour l'anglais, la première révolution du domaine du traitement automatique des langues, celle de l'émergence des *approches statistiques* (Hall *et al.*, 2008 ; Church, 2011). Ces approches reposent

sur la disponibilité de jeux de données de volume important et ont conduit à des progrès importants sur de nombreuses tâches et pour de nombreuses langues (cf. par exemple, pour la traduction automatique, (Brown *et al.*, 1990, 1993) [1]). Cette révolution n'a toutefois touché dans un premier temps que l'anglais. Elle n'a atteint d'autres langues que progressivement, au cours des années 2000 voire 2010, et n'est toujours pas pleinement d'actualité pour les langues les moins dotées. Les approches statistiques relèvent de l'apprentissage automatique et, de ce fait, ont changé la façon dont l'expertise linguistique est encodée et fournie aux systèmes computationnels. Les approches par règles, ou *approches symboliques*, encodaient cette expertise dans des règles (la grammaire) et des ressources lexicales. Les approches statistiques, souvent supervisées, s'appuient sur des données annotées dont des informations linguistiques peuvent être inférées afin d'en extraire les régularités et généralisations qui y sont implicites. Mais de telles données annotées restent de taille limitée, et la distribution zipfienne des mots résulte en leur insuffisance pour couvrir l'ensemble des faits lexicaux. Les ressources lexicales gardent ainsi toute leur importance dans les systèmes statistiques en tant que moyen de faire face à ce phénomène de *dispersion lexicale* (*lexical sparsity* en anglais), comme cela a été montré sur différentes tâches comme l'étiquetage morphosyntaxique (Goldberg *et al.*, 2009), l'analyse syntaxique (Collins, 1997 ; Riezler *et al.*, 2002 ; Versley et Rehbein, 2009) ou la traduction automatique (Carpuat et Wu, 2007). Les approches symboliques conservent également toute leur pertinence, surtout lorsqu'elles sont complétées plutôt que remplacées par des modèles statistiques. Ainsi, au niveau syntaxique, les modèles formels faiblement contextuels, que certains ont pu considérer comme inutiles suite à l'arrivée de modèles statistiques non contextuels, ont repris de l'importance, couplés avec des architectures statistiques (Villemonte de La Clergerie, 2013).

Au milieu des années 2010, une autre révolution est apparue dans ce contexte, la révolution neuronale (Manning, 2015 ; Goldberg, 2017). Tout comme la révolution statistique, qui trouvait ses racines dans des théories anciennes et des travaux déjà classiques, notamment en reconnaissance de la parole, les *approches neuronales* s'appuient sur des techniques et des résultats vieux de plusieurs décennies. Elles relèvent elles aussi de l'apprentissage automatique. Elles ont été rendues possibles par la croissance exponentielle de la production de données brutes, par l'augmentation du volume, de la richesse et de la variété des données annotées, tant en termes de niveaux linguistiques couverts qu'au niveau des langues concernées, mais aussi par la poursuite de l'augmentation de la puissance de calcul des systèmes informatiques, en particulier avec la réapparition de co-processeurs dédiés aux calculs numériques, les processeurs graphiques (ou GPU). Cette puissance de calcul et ce volume de données ont permis l'émergence de

1. Leur acceptation n'a pas été immédiate, mais le second de ces deux articles a été récemment primé par un comité lié à la conférence NAACL 2018 comme l'un des trois articles qui a le mieux résisté à l'épreuve du temps (cf. `https://naacl2018.wordpress.com/2018/03/22/test-of-time-award-papers/`).

systèmes utilisant des réseaux de neurones d'une complexité sans précédent, conduisant à ce que l'on qualifie désormais d'*apprentissage profond* (ou du terme anglais *deep learning*). Ces approches, qui ont permis à des domaines comme la vision par ordinateur ou le traitement de la parole de faire des progrès considérables, ont également amélioré l'état de l'art dans la plupart des tâches relevant du traitement automatique des langues, quoique dans des proportions variables. Certains de ces progrès résultent toutefois d'idées pré-existantes. Ainsi, l'idée consistant à représenter les mots sous forme de vecteurs dans des espaces continus (plongements lexicaux, ou *word embeddings*; cf. Mikolov *et al.* (2013)) a été mise en œuvre de différentes manières bien avant l'émergence des approches neuronales dans notre domaine (cf. par exemple Landauer *et al.*, 1998). Une telle façon de représenter les mots est néanmoins centrale dans le paradigme neuronal, et constitue, grâce à l'exploitation de volumes importants de données non annotées, une autre réponse au problème de la dispersion des données. Reste à savoir si ressources lexicales et plongements lexicaux sont des réponses complémentaires ou redondantes à ce problème.

Au cours de l'évolution de cette discipline qu'est le traitement automatique des langues, son rapport à la linguistique a évolué. Les approches symboliques reposaient de manière cruciale sur les modèles issus des travaux en linguistique. Elles ont soulevé des questions linguistiques pertinentes et difficiles et ont contribué à faire avancer les recherches dans ce domaine. Elles ont induit le développement de ressources linguistiques, et notamment lexicales, qui ont également contribué à une meilleure description et compréhension des langues**TODO : refs**. Elles ont enfin permis l'identification et la modélisation de généralisations linguistiques, au travers des propriétés de modèles formels et computationnels. Or l'émergence des approches statistiques et plus encore neuronales a parfois fait penser que les travaux en linguistique n'étaient plus utiles. Après tout, on peut entraîner un analyseur syntaxique statistique en quelques heures, alors qu'il faut des années pour développer une grammaire informatique à large couverture. Mais c'est oublier que l'analyseur statistique est entraîné sur des données d'apprentissage, dont le développement prend lui aussi des années et repose sur des choix d'analyse linguistique qui reflètent des positions théoriques. Simplement, ce ne sont pas toujours les mêmes informations linguistiques qui sont encodées dans les deux cas, et, plus important encore, elles ne le sont pas de la même façon. De même, les informations linguistiques encodées dans un lexique ne sont pas les mêmes que celles encodées dans un jeu de données annotées. Le lexique pourra contenir des informations plus riches, y compris sur des mots plus rares et des constructions rarement attestées dans les jeux de données. Il reste à comprendre comment formaliser ces informations, comment développer des ressources lexicales, et comment exploiter au mieux ces informations dans les systèmes de traitement automatique.

Ainsi, la place de la linguistique par rapport au traitement automatique des langues ne doit pas être sous-estimée. Le succès des approches statistiques et neuronales ne doit pas faire oublier que le traitement automatique des langues s'est pensé comme un domaine d'application de la linguistique, et ne doit pas non plus conduire à l'abandon de ce lien indispensable entre science des données et science de leur traitement, surtout sur des données aussi complexes et variées que les données langagières. Néanmoins, les progrès en traitement automatique des langues ont également permis, sans qu'il n'y ait ici de contradiction, de renverser la situation. En effet, les approches quantitatives, et notamment computationnelles, de l'étude de questions linguistiques se sont multipliées, en particulier au cours de la dernière décennie. Ces approches, qui relèvent de la *linguistique computationnelle*, permettent d'investiguer par des moyens informatiques un certain nombre de questions linguistiques nouvelles, ou d'apporter à des questions anciennes des réponses renouvelées. En un sens, la linguistique est ainsi devenue un domaine d'application du traitement automatique des langues.

Mon travail se positionne ainsi à l'interface entre informatique et linguistique, entre traitement automatique des langues et linguistique computationnelle. Je pense qu'une meilleure connaissance et une meilleure compréhension de la langue et des langues doivent pouvoir améliorer les systèmes de traitement automatique. Je pense également que de tels systèmes peuvent contribuer à une meilleure connaissance et une meilleure compréhension de la langue et des langues. Et dans cet aller-retour entre linguistique et informatique, je considère le lexique comme un enjeu crucial. C'est ce qui explique qu'une partie significative de mes travaux, depuis ma thèse, concerne directement ou indirectement les informations lexicales.

Structure du document

Le présent document présente un ensemble de travaux que j'ai menés depuis une dizaine d'années en traitement automatique des langues et en linguistique computationnelle. La thématique générale que j'ai retenue ici est celle des ressources lexicales, thématique à laquelle se rattache directement ou indirectement une partie importante de mes activités de recherche. J'ai ainsi étudié des problématiques associées à la modélisation des informations lexicales, au développement de ressources lexicales et à l'utilisation de telles ressources, à la fois dans des architectures de traitement automatique des langues et pour étudier des questions plus linguistiques. Je me suis concentré sur trois niveaux d'analyse qui jouent un rôle-clef en traitement automatique des langues : les niveaux morphologique, syntaxique et sémantique.

J'ai réalisé ces travaux en tant que membre de l'équipe-projet Inria et UMR-I ALPAGE (Inria et Université Paris–Diderot) jusqu'en 2016, puis de l'équipe Inria qui lui fait suite,

ALMAnaCH (Inria et École Pratique des Hautes Études) [2]. J'ai été responsable d'ALPAGE de début 2014 à son terme fin 2016, et suis actuellement responsable d'ALMAnaCH.

Avant de traiter d'unités lexicales et de lexiques, il m'a semblé nécessaire que le premier chapitre soit le lieu d'une réflexion sur ces deux notions difficiles à définir. Ce chapitre ne saurait constituer une étude approfondie de l'insaisissable notion de « mot », qui dépasserait très largement le cadre de ce document. Mais il apporte quelques éléments et quelques définitions qui seront utilisées dans les chapitres suivants. Le point de vue que j'adopte consiste à faire l'hypothèse qu'une production langagière peut être segmentée en unités élémentaires, et à étudier les différentes propriétés que l'on peut vouloir associer à de telles unités. Sans surprise, les différents types de propriétés ne correspondent pas nécessairement aux mêmes unités élémentaires, et donc pas nécessairement à la même segmentation. Cela permet de justifier la définition de plusieurs types de « mots », et donc d'unités lexicales et de lexiques, en cohérence avec la variété des travaux qui font l'objet des chapitres ultérieurs. Je distingue ainsi, entre autres, les notions de *mot typographique* (ou *token*), de *(mot-)forme*, de *mot morphosyntaxique*, de *mot syntaxique* (ou *mot grammatical*), et de *mot sémantique*.

Le chapitre 2 est le premier des trois chapitres consacrés au niveau morphologique. Il est consacré à mes travaux sur la formalisation des informations morphologiques flexionnelles, tant du côté des informations lexicales que du côté de la grammaire morphologique qui les accompagne dans un lexique morphologique. J'y décris la façon dont les informations morphologiques peuvent être encodées dans l'architecture Alexina, laquelle rassemble tout à la fois un formalisme de description des informations lexico-morphologiques et lexico-syntaxiques, un ensemble d'outils pour l'acquisition d'informations lexicales et la gestion de lexiques, et une collection de lexiques morphologiques et parfois syntaxiques pour un certain nombre de langues variées. Alexina implémente en réalité deux formalismes morphologiques qui, bien qu'ils partagent un certain nombre de propriétés, diffèrent sur d'autres plans. Le premier est le formalisme d'origine d'Alexina, sur lequel s'appuie notamment le lexique Alexina le plus avancé et le plus utilisé, le Le*fff*, lexique morphologique et syntaxique du français. Le second est une adaptation et une extension du formalisme morphologique PARSLI développé par Walther (2013b, 2016), qui est donc brièvement présenté. L'annexe B fournit quelques détails sur l'implémentation Alexina de PARSLI, illustrés par des exemples issus des lexiques MaltLex et Le*ff*la, lexiques Alexina du maltais (sémitique, afro-asiatique, Malte) et du latin (italique, indo-européen).

Le chapitre 3 décrit certains travaux que j'ai menés en vue du développement de lexiques morphologiques, ainsi que certaines techniques automatiques que j'ai été amené à développer à cette fin, destinées à accélérer le processus et à limiter le travail manuel.

2. `http://team.inria.fr/almanach/fr`

Je décris dans un premier temps quatre approches pour le développement de lexiques flexionnels, qui se distinguent par la nature des informations disponibles en entrée : corpus brut et grammaire morphologique, lexique de formes fléchies, données lexicales semi-structurées extraites de ressources wiki, et finalement lexique morphologique pour une langue typologiquement proche. Je présente ensuite les travaux réalisés dans le cadre du projet ANR EDyLex, dont j'étais le porteur, sur l'extension de lexiques morphologiques à partir de corpus dynamiques (flux de dépêches d'agence) par l'analyse de la structure morphologique (dérivationnelle et compositionnelle) de candidats néologismes. Enfin, je présente deux expériences d'acquisition automatique de liens dérivationnels entre entrées lexicales flexionnelles, l'une à partir de règles de dérivation explicites et d'un corpus brut, l'autre de façon endogène au sein d'un lexique flexionnel.

Le chapitre 4 rassemble plusieurs travaux en morphologie quantitative que j'ai réalisés à partir de lexiques morphologiques. Ces travaux s'articulent autour de la notion de complexité morphologique. Je présente tout d'abord une mesure de la complexité d'un système flexionnel décrit sous la forme d'un lexique Alexina et qui s'appuie sur la notion de longueur de description, au sens de la théorie de l'information. Cette mesure est ensuite utilisée pour comparer des analyses concurrentes de systèmes morphologiques. Sur le système verbal du français, elle nous a permis d'étudier la façon dont l'information morphologique pouvait être distribuée entre la grammaire morphologique et le lexique proprement dit. Sur le khaling (kiranti, sino-tibétain, Népal), elle nous a permis d'étudier la pertinence de l'utilisation de traits abstraits, dits *morphomiques* (cf. Aronoff, 1994). Un inconvénient de ce type de mesures, que l'on peut qualifier de *constructives* au sens de Blevins (2006), est qu'elles s'appliquent à des descriptions formelles en partie arbitraires du système morphologique et non aux paradigmes flexionnels eux-mêmes. J'ai donc étudié une autre famille de mesures, *abstractives* au sens de Blevins (2006), qui s'appuient directement sur les inventaires de formes fléchies et sur la notion d'entropie conditionnelle entre cases des paradigmes. Après avoir identifié et critiqué la métrique proposée à cet égard par Ackerman *et al.* (2009), je propose une nouvelle métrique qui évite certaines de ses difficultés. Mais certains problèmes demeurent. Notamment, aucune de ces métriques abstractives ne prend en compte le caractère structuré de l'ensemble des comportements flexionnels d'un système morphologique, c'est-à-dire, pour simplifier, la notion de classe flexionnelle. Je montre alors qu'il est possible de faire émerger une hiérarchie de comportements flexionnels par une approche couplant un point de vue abstractif sur les paradigmes de formes fléchies et un critère plus constructif, qui reprend la notion de longueur de description, appliqué à des ensembles de comportements flexionnels. Des expériences sur les systèmes verbaux du français et du portugais permettent alors de faire émerger uniquement à partir des paradigmes flexionnels ce que nous avons appelé des *micro-classes*, c'est-à-dire des comportements

flexionnels totalement spécifiés, et des *macro-classes*, qui correspondent pour partie aux classes flexionnelles des grammaires traditionnelles.

Avec le chapitre 5, nous quittons le niveau morphologique pour le niveau syntaxique, celui des cadres de sous-catégorisation (valence), des diathèses (passif, impersonnel…) et des relations relevant de la syntaxe profonde (phénomènes de contrôle, d'attribution…). La première partie de ce chapitre traite de la modélisation de l'information lexico-syntaxique. Après une brève description des lexiques syntaxiques qui existaient pour le français au début des années 2000, je décris leurs limites, et notamment les limites des modèles sous-jacents. J'explique ainsi ce qui m'a motivé à étendre le L*efff* au niveau syntaxique et à développer à cette fin un nouveau modèle de représentation des propriétés lexico-syntaxiques, modèle qui constitue le niveau syntaxique du formalisme Alexina. Je décris ensuite un certain nombre de techniques semi-automatiques qui nous ont permis d'accélérer le développement du L*efff*, notamment grâce à l'interprétation et à l'exploitation d'autres ressources lexicales, mais également au moyen d'une technique de fouille d'erreurs dans les sorties d'un analyseur syntaxique s'appuyant sur le L*efff*. Le chapitre se termine par quelques éléments d'évaluation du L*efff*.

Le chapitre 6 relève du niveau sémantique, puisqu'il décrit le travail que j'ai mené dans le cadre du développement de deux wordnets, le WOLF, « WOrdnet Libre du Français », et sloWNet, wordnet du slovène (slave, indo-européen, Slovénie). La première étape du développement de ces wordnets est totalement automatique et s'appuie à la fois sur des corpus parallèles multilingues alignés et sur des ressources lexicales bilingues. Une seconde étape a permis une augmentation importante de la couverture de ces deux wordnets grâce à une meilleure exploitation d'un plus grand nombre de ressources lexicales bilingues. Plusieurs approches que j'évoque brièvement ont alors permis d'étendre le WOLF, avant une troisième étape importante dans le développement du WOLF et de sloWNet, qui s'appuie sur une technique d'identification d'erreurs potentielles dans un wordnet. Le WOLF, qui nécessiterait encore un travail de validation et de nettoyage pour être vraiment fiable, n'en est pas moins le premier wordnet libre pour le français, et le plus grand wordnet pour cette langue ; sloWNet, de son côté, a bénéficié d'un effort plus important de validation manuelle, et constitue aujourd'hui une ressource fiable, le seul wordnet à grande échelle pour le slovène.

Les trois chapitres qui suivent traitent de l'exploitation d'informations lexicales dans des systèmes de traitement automatique, successivement pour l'analyse de surface, l'étiquetage morphosyntaxique et l'analyse syntaxique. Le chapitre 7 s'attaque à la problématique du traitement de corpus bruts, éventuellement bruités, en partie dans le cadre de la chaîne de traitement SxPipe. Après une discussion sur les étapes de segmentation en tokens et en phrases, je décris deux ensembles de travaux qui traitent de l'identification de mots grammaticaux dans une séquence de tokens. Le premier

ensemble regroupe différents travaux en correction orthographique, déterministe ou non, qui s'appuient sur des règles développées manuellement ou acquises automatiquement sur corpus au moyen d'une approche par analogie. Le deuxième ensemble de travaux concerne la problématique de la segmentation non supervisée du mandarin, langue dont le système d'écriture ne fait pas usage de séparateurs typographiques. Les techniques utilisées s'appuient à la fois sur la notion d'entropie (en l'espèce, l'entropie de branchement) et sur la notion de longueur de description, faisant ainsi indirectement écho à mes travaux en morphologie quantitative décrits au chapitre 4.

Le chapitre 8 traite de l'exploitation d'informations lexicales pour l'étiquetage morphosyntaxique. Je présente tout d'abord trois sytèmes d'étiquetage morphosyntaxiques statistiques et neuronaux que j'ai développés pour expérimenter cette problématique. Je décris tout d'abord MElt, système reposant sur les modèles de Markov à maximisation d'entropie, qui peut utiliser un lexique externe comme source de traits supplémentaires. Je m'attarde sur le cas du français, sur lequel MElt est resté plusieurs années au niveau de l'état de l'art, et montre notamment l'impact relatif de la taille du corpus d'entraînement et celle du lexique externe sur les performances de l'étiqueteur. Je présente ensuite des expériences multilingues impliquant MElt, ainsi qu'un système d'encapsulage (*wrapping*) de MElt par un reconnaisseur d'entités nommées afin d'améliorer l'étiquetage de ces dernières. Je présente ensuite alVWtagger, étiqueteur statistique qui fait suite à MElt et qui, grâce à des lexiques morphologiques extraits de façon *ad hoc*, m'a permis d'être classé 3$^{\text{ème}}$ sur 33 en étiquetage en parties du discours lors de la campagne CoNLL 2017 sur l'analyse syntaxique de corpus bruts en *Universal Dependencies* (Zeman *et al.*, 2017). Je décris enfin alNNtagger, étiqueteur neuronal qui étend le système de Plank *et al.* (2016) pour permettre l'exploitation d'informations lexicales externes, et grâce auquel j'ai montré que de tels lexiques permettaient d'améliorer les performances, quand bien même sont également utilisés des plongements lexicaux (*word embeddings*). Ce chapitre se termine par la description de l'approche que j'ai développée pour l'étiquetage morphosyntaxique de corpus bruités issus du web, et de son application dans deux contextes. Le premier d'entre eux est le développement du *French Social Media Bank*, corpus arboré de données produites par les utilisateurs de réseaux sociaux et de forums en ligne. Le second contexte d'utilisation est la campagne d'évaluation SANCL 2012 sur l'analyse syntaxique de corpus anglais issus du web (Petrov et McDonald, 2012).

Le chapitre 9 est consacré à l'apport des lexiques morphologiques et syntaxiques pour l'analyse syntaxique. Dans une première section, je décris tout d'abord différentes expériences réalisées avec l'analyseur hybride FRMG du français (Villemonte de La Clergerie, 2014), qui, grâce à une métagrammaire développée manuellement, aux niveaux morphologiques et syntaxiques du Le*fff* et à des mécanismes de désambiguïsation

statistiques appris automatiquement, a de très bonnes performances [3]. Je montre en particulier que remplacer le Le*fff* par d'autres ressources lexicales syntaxiques converties au format Alexina conduit à une dégradation des performances. Je mentionne également la façon dont j'ai adapté la technique de fouille d'erreurs dans les sorties d'analyseurs syntaxiques, technique mentionnée ci-dessus, afin d'identifier les points forts et les points faibles des différents lexiques syntaxiques ainsi comparés. La seconde section est consacrée à l'exploitation de lexiques morphologiques dans des analyseurs syntaxiques statistiques en constituants. Je décris l'une des expériences que j'ai menées dans cette direction, à savoir une expérience sur l'espagnol s'appuyant sur le Le*ff*e, lexique Alexina pour cette langue, et sur l'architecture d'analyse syntaxique LORG (Le Roux *et al.*, 2012b). L'analyseur ainsi construit était alors au niveau de l'état de l'art. La troisème section de ce chapitre concerne l'analyse syntaxique en dépendances, et se penche cette fois-ci sur l'exploitation du niveau syntaxique du Le*fff*, afin d'améliorer la qualité des cadres de sous-catégorisation produits par l'analyseur MATE (Bohnet, 2010). L'apport des cadres de sous-catégorisation fournis par le Le*fff* est comparé à celui de cadres extraits d'un corpus analysé automatiquement et de cadres extraits d'un corpus arboré. Enfin, ce chapitre se termine par la description de travaux sur l'analyse syntaxique de corpus bruités, lesquels, comme au chapitre précédent, concernent successivement le développement du *French Social Media Bank* et notre participation à la campagne d'évaluation SANCL 2012, à laquelle nous avons été classés seconds.

Ce document se termine au chapitre 10 par un bref bilan des travaux présentés, des évolutions actuelles du domaine du traitement automatique des langues, des directions de recherche que j'envisage pour l'équipe dont je suis le responsable, et, pour finir, sur une direction de recherche nouvelle que je souhaite explorer dans les prochaines années : celle de l'étymologie computationnelle, c'est-à-dire la modélisation de l'évolution lexicale des langues, et notamment des langues indo-européennes anciennes, en prenant en compte les dimensions phonétiques, phonologiques, morphologiques et sémantiques, tout en traitant des multiples mécanismes de création lexicale et d'emprunt.

Pour finir cette section, il est important de signaler que pour éviter que ce document, déjà long, ne prenne une ampleur trop excessive, j'ai décidé de réserver à une annexe des aperçus de l'histoire de la recherche dans les domaines ici abordés. On trouvera ainsi successivement dans l'annexe A :

— un aperçu des approches formelles de la morphologie,

— un état de l'art sur le développement de lexiques flexionnels puis dérivationnels,

— une discussion sur la quantification et la mesure de la complexité morphologique,

3. Un état de l'art des analyseurs syntaxiques pour le français à l'époque de ces travaux est fourni en annexe à la section A.10.3.

— un bref panorama des modèles utilisés pour la représentation d'informations lexico-syntaxiques suivi d'un historique des travaux concernant le développement de lexiques syntaxiques,

— un état de l'art des techniques automatiques de développement de wordnets,

— un très bref aperçu des travaux académiques en correction orthographique,

— un historique des travaux en étiquetage morphosyntaxique,

— et pour finir un historique des travaux en analyse syntaxique.

Dans le corps de ce document, il sera fait référence à cette annexe historique en tant que de besoin.

Encadrement et collaborations

Nombre de mes travaux ont été réalisés en collaboration. Je ne citerai ici que les doctorants avec lesquels j'ai collaboré, y compris mais pas seulement en tant qu'encadrant principal de quatre thèses de doctorat, ainsi que les post-doctorants placés sous ma responsabilité dans le cadre de divers projets financés. Mes autres collaborateurs sont mentionnés au travers des références à nos articles communs.

J'ai été successivement l'encadrant principal de quatre thèses de doctorat :

— Rosa Stern, sur la détection et le liage d'entités nommées dans des dépêches d'agence (cf. ci-dessous ; thèse CIFRE en partenariat avec l'Agence France-Presse et réalisée dans le contexte du projet ANR EDyLex, dont j'étais le porteur (2009–2012, enrichissement dynamique de lexiques) ; directrice officielle : Laurence Danlos ; voir notamment la sections 3.2.3, la note de bas de page 21 du chapitre 7, mais également la section suivante) ;

— Valérie Hanoka, sur l'extraction et l'extension de terminologies multilingues (cf. ci-dessous ; thèse CIFRE en partenariat avec Verbatim Analysis, start-up Inria dont je suis l'un des fondateurs ; directrice officielle : Laurence Danlos ; voir la section suivante ainsi que la mention qui est faite d'un des aspects de ce travail à la section 6.2) ;

— Pierre Magistry, sur la segmentation non supervisée du chinois mandarin (co-directeurs : Sylvain Kahane et Marie-Claude Paris ; voir notamment la section 7.4).

— Marion Baranes, sur la correction orthographique de commentaires clients, y compris via l'identification de néologismes à ne pas corriger (thèse en partenariat avec l'entreprise viavoo ; directrice officielle : Laurence Danlos ; voir notamment les sections 3.2.3, 3.3.2 et 7.3.2).

Je deviendrai dans les prochains mois le directeur de thèse ou l'un des co-encadrants de deux ou trois nouvelles thèses selon le devenir d'une demande de financement, dont

l'une, dont le financement est confirmé, sous forme de convention CIFRE avec le Facebook Artificial Intelligence Research (FAIR).

J'ai également collaboré avec d'autres personnes préparant alors un doctorat, collabortions qui se sont parfois poursuivies après leur soutenance, et notamment :

— Elsa Tolone, sur les ressources lexicales syntaxiques et leur utilisation pour l'analyse syntaxique (cf. notamment la section 9.1) ;

— Lionel Nicolas, sur la correction semi-automatique de lexiques syntaxiques (cf. la mention qui en est faite à la section 5.2.2) et sur le développement de ressources lexicales pour l'espagnol et le galicien ;

— Darja Fišer, sur le développement de lexiques sémantiques de type wordnet (cf. chapitre 6) ;

— Géraldine Walther, en morphologie formelle et quantitative ainsi qu'en développement de ressources lexicales et d'étiqueteurs morphologiques pour des langues peu dotées (persan, variétés du kurde, khaling, romanche…) ;

— Jana Strnadová, sur le développement de lexiques dérivationnels (cf. section 3.3.1) ;

— Sacha Beniamine, en morphologie quantitative (cf. section 4.3).

Enfin, j'ai travaillé avec cinq post-doctorants financés par des projets dont j'étais porteur (ANR EDyLex, axe 6 « Ressources Linguistiques » du LabEx *Empirical Foundations of Linguistics*) ou responsable local (projets FUI PACTE, coordonné par l'entreprise Numen Digital, et ANR RAPID VerDI, coordonné par la start-up Trooclick/Storyzy).

— Marianna Apidianaki (ANR EDyLex), sur l'enrichissement de lexiques sémantiques de type wordnet (travaux mentionnés mais non détaillés à la section 6.2) ;

— Damien Nouvel (ANR EDyLex), sur l'enrichissement de lexiques morphologiques à partir de flux textuels (cf. section 3.2) ;

— Kata Gábor (ANR EDyLex puis FUI PACTE), sur l'enrichissement de lexiques sémantiques de types wordnet (travaux mentionnés mais non détaillés à la section 6.2) et la correction orthographique post-OCR, notamment au sein de mentions d'entités nommées spécialisées (cf. sections 7.2.3 et 7.3.3) ;

— Yves Scherrer (LabEx *Empirical Foundations of Linguistics*), sur l'acquisition automatique de lexiques flexionnels à partir d'un lexique flexionnel pour une langue étymologiquement proche (cf. section 3.1.4) ;

— Héctor Martínez Alonso (ANR RAPID VerDI), sur la détection d'omissions dans des contenus journalistiques (travaux non mentionnés dans ce document) et sur l'étiquetage morphosyntaxique neuronal (cf. section 8.3).

Travaux de recherche non évoqués dans ce document

Depuis que j'ai soutenu ma thèse de doctorat en 2006, j'ai travaillé sur une grande variété de sujets en traitement automatique des langues et en linguistique computationnelle. Certains de mes travaux sont décrits avec plus ou moins de détails dans les chapitres qui suivent. D'autres sont simplement évoqués, parfois seulement au détour d'une note de bas de page. Mais certains ne sont pas cités dans ce document, soit par choix, soit parce qu'ils sont trop récents, soit parce qu'ils ne se seraient pas intégrés, ou difficilement, dans la thématique générale de la modélisation, l'acquisition et l'exploitation des informations lexicales. Je me permets donc ici d'en faire un inventaire presque exhaustif.

Un premier travail qui aurait pu avoir sa place dans ce document, et qui prépare certains travaux futurs évoqués au chapitre 10, a consisté en la construction automatique d'une base de données étymologiques à partir du *wiktionary* (Sagot, 2017a,b). Ce travail est indirectement en lien avec des recherches que j'ai effectuées en linguistique historique *stricto sensu* en collaboration avec Romain Garnier et qui portent sur les langues indo-européennes. Certaines se situent au niveau proto-indo-européen (Garnier *et al.*, 2017 ; Garnier et Sagot, 2018a ; Sagot, 2018a), d'autres reflètent ma prédilection pour le grec ancien et les substrats et adstrats qui ont contribué à construire son lexique (Garnier et Sagot, 2015, 2017, 2018b).

Un pan important de mes travaux à l'interface entre ressources lexicales et analyse de surface n'est mentionné qu'en passant dans ce document. Il s'agit de mes travaux sur la reconnaissance et le liage des entités nommées. Ces travaux, qui se sont déroulés en partie dans le contexte du projet FUI Scribo puis du projet ANR EDyLex mentionné ci-dessus, ont été réalisés en grande majorité dans le cadre de la thèse de Rosa Stern. Comme indiqué ci-dessus, il s'agissait d'une thèse en partenariat avec l'Agence France-Presse (AFP) et dont j'étais le principal encadrant. Ces travaux ont donné lieu à plusieurs publications, notamment concernant le développement de ressources linguistiques telles que la base d'entités Aleda (Stern et Sagot, 2010b ; Sagot et Stern, 2012) et l'annotation référentielle du Corpus Arboré de Paris 7 en entités nommées (Sagot *et al.*, 2012), mais également des travaux en détection et liage d'entités nommées par des techniques couplant règles, heuristiques et modèles statistiques (Stern et Sagot, 2010b,a ; Béchet *et al.*, 2011 ; Stern *et al.*, 2012 ; Stern et Sagot, 2012).

En lien avec cette thématique, par exemple dans le cadre du projet Scribo, j'ai également travaillé sur les citations dans les dépêches d'agence, là encore en partenariat avec l'AFP. Ces travaux ont couvert des aspects linguistiques (Danlos *et al.*, 2010 ; Sagot et Danlos, 2010), des aspects liés au développement de ressources lexicales dédiées (Sagot *et al.*, 2010) et des aspects informatiques, pour la détection et l'attribution de citations dans des dépêches (Villemonte de La Clergerie *et al.*, 2009b), qui s'appuyaient sur les précédents.

Ces travaux sont allés de pair avec mes travaux sur les entités nommées, et ont conduit à une application web sur le site du journal *Libération* en partenariat avec l'AFP à l'occasion des élections présidentielles de 2012, ainsi qu'à une application sur *smartphone* publiée par l'AFP elle-même. Une démonstration désormais obsolète de ces technologies sur des dépêches AFP en lien avec ces élections est toujours accessible en ligne [4].

Parmi mes autres travaux en lien avec le développement de ressources lexicales, on peut citer ma participation au projet ANR ASFALDA sur le développement d'un FrameNet du français (Candito *et al.*, 2014) et ma participation à des réflexions sur la pertinence et la légitimité des approches faisant usage de la myriadisation du travail parcellisé (*crowdsourcing*) en traitement automatique des langues (Adda *et al.*, 2011 ; Sagot *et al.*, 2011a ; Fort *et al.*, 2014).

Mes travaux en grammaires formelles et en analyse syntaxique symbolique ne sont pas mentionnés non plus dans ce document. Sur ces sujets, j'ai notamment travaillé sur les aspects formels et algorithmiques de l'analyse syntaxique non contextuelle (Boullier *et al.*, 2009 ; Boullier et Sagot, 2009b, 2010) et sur des problématiques relevant directement de la théorie des grammaires formelles (Boullier et Sagot, 2009a ; Sagot et Satta, 2010). Est également absent de ce document mon travail sur l'analyseur statistique faiblement contextuel MICA (Bangalore *et al.*, 2009), ni ma contribution récente à la campagne d'évaluation orientée-tâche des analyseurs syntaxiques (Schuster *et al.*, 2017).

Dans document, je fais également l'impasse sur mes travaux en collaboration avec Djamé Seddah sur la modélisation formelle des coordinations elliptiques, à l'interface entre linguistique formelle et grammaires formelles (Seddah et Sagot, 2006b,a ; Seddah *et al.*, 2010b).

Enfin, je ne mentionne pas non plus mes travaux concernant l'analyse automatique d'enquêtes internes (enquêtes réalisées auprès des employés d'une entreprise). Ces travaux ont pourtant occupé une part significative de mon temps depuis 2009, date où j'ai créé la start-up Verbatim Analysis avec Dimitri Tcherniak, et même en vérité depuis plusieurs années auparavant. Plus récemment, j'ai créé une nouvelle start-up dans ce domaine, opensquare, dans laquelle je travaille notamment avec Dimitri Tcherniak et Olivier Hamelle. On pourra noter que la thèse de Valérie Hanoka, mentionnée ci-dessus, a été co-financée par Verbatim Analysis dans le cadre d'une convention CIFRE, avec plusieurs publications à la clef, indiquées au chapitre 6. Une expérience ponctuelle réalisée au sein de Verbatim Analysis a donné lieu à une publication dans une conférence spécialisée en psychosociologie des organisations (Čingienė *et al.*, 2015).

4. `http://alpage.inria.fr/sapiens/index.html`.

Ressources lexicales libres

Pour terminer cette introduction, il me semble important d'exprimer ma position sur une problématique importante, celle des modalités de distribution des ressources lexicales. Il s'agit là en effet d'une question dont les enjeux sont multiples (politiques, juridiques, éthiques) et à laquelle les réponses que l'on apporte ont un impact sur l'avancée de l'état de l'art.

Le coût élevé du développement de ressources lexicales reste une réalité, malgré l'utilisation de techniques automatiques. En effet, une ressource n'est souvent vraiment utilisable qu'après un travail de validation manuelle au moins partielle. Dans tous les cas, un tel travail reste sujet à de nombreuses erreurs. Ceci provient de ce qu'une ressource lexicale à large couverture peut rassembler une quantité importante d'informations complexes sur un nombre considérable d'unités linguistiques. De plus, la plupart des ressources décrivent le lexique d'une langue donnée, et développer une ressource équivalente pour une autre langue nécessite généralement un effort important. C'est là l'une des raisons de l'importance croissante que prennent les ressources librement disponibles : les auteurs et les propriétaires, tout en conservant respectivement leurs droits moraux et leurs droits patrimoniaux, permettent l'exploitation, la transformation et la redistribution d'une telle ressource, sous réserve que les clauses spécifiées dans la licence qui lui est associée soient respectées.

L'accès libre est le meilleur moyen dont disposent les développeurs de ressources lexicales et leurs utilisateurs pour surmonter les trois défis principaux que sont les suivants.

1. La visibilité : il s'agit d'une condition préalable à la disponibilité immédiate et pérenne, une ressource qui n'est pas visible ne pouvant pas être trouvée par ses utilisateurs potentiels. Cette problématique, souvent sous-évaluée, peut être en partie résolue grâce aux structures de partage de métadonnées comme la plateforme européenne META-SHARE.
2. La disponibilité immédiate : c'est une condition nécessaire à ce qu'une ressource soit citée, utilisée et par conséquent améliorée ; c'est également indispensable pour garantir la reproductibilité des résultats scientifiques obtenus grâce à la ressource.
3. La disponibilité pérenne : les informations contenues dans les ressources librement disponibles ont plus de chances d'être utilisables sur le long terme, et notamment après que leurs développeurs ont cessé d'y travailler, puisque de telles ressources peuvent être redistribuées par autrui sous leur forme d'origine, après modifications ou après intégration à d'autres ressources. Ceci montre également qu'il est dans l'intérêt des agences de financement de promouvoir au sein des projets qu'elles financent le développement de ressources libres, qui ont bien plus de chances de

pouvoir être réutilisées dans d'autres projets, dans d'autres contextes, y compris des années plus tard.

L'ensemble des ressources que nous avons développé et dont il sera question dans ce document sont ainsi librement disponibles, tout comme les outils de développement et de manipulation qui leur sont associés.

Chapitre 1

Modéliser l'unité lexicale

Sommaire

La notion de *lexique* joue un rôle central dans les travaux présentés dans ce document. Il convient donc d'en proposer une définition. Intuitivement, *a minima* dans le domaine du traitement automatique des langues et de la linguistique computationnelle, un lexique est une collection de « mots » associés à des propriétés linguistiques. Mais une telle définition n'a de sens que si l'on définit au préalable la notion de « mot » et celle de *propriété linguistique*. Nous verrons notamment que la notion de mot est particulièrement rétive à toute définition objective et satisfaisante [1]. Néanmoins, travailler, comme nous l'avons fait,

1. Haspelmath (2011, p. 24), dans sa publication de référence sur le sujet, en propose trois illustrations sous la forme de citations :

— Jespersen (1924, p. 92) : « Qu'est-ce qu'un mot ? et qu'est qu'un seul mot ? Ce sont là des problèmes très difficiles… » (*What is a word ? and what is one word ? These are very difficult problems…*) ;

à la représentation et à l'exploitation informatique des informations relevant du lexique des langues impose de savoir de quoi l'on parle, et donc de réfléchir à la façon dont on peut définir correctement les notions que l'on est amené à manipuler.

1.1 Lexique et unités lexicales

La notion même de lexique, qui correspond intuitivement à une collection de « mots » associés à des propriétés linguistiques, repose sur deux hypothèses qui sont des simplifications motivées par le contexte qui est le nôtre, celui du traitement automatique de données textuelles :

- il est raisonnable de représenter des données langagières, qui sont fondamentalement des signaux sonores [2], sous forme de séquences d'unités minimales ; il y a donc deux niveaux de segmentation, tous deux discutables d'un point de vue théorique mais néanmoins utiles en pratique : un niveau macroscopique qui correspond à la notion d'*énoncé* (ou *phrase*) et un niveau microscopique, celui des unités minimales qui, dans notre cas, puisque nous traitons de données écrites, sont les graphèmes ;
- ces énoncés peuvent être segmentés en *unités élémentaires* qui sont des séquences (contigues ou non) d'unités minimales ; ces unités élémentaires peuvent faire l'objet de généralisations linguistiques.

De telles généralisations consistent en l'identification de différents types de propriétés des ces unités élémentaires, qui relèvent de différents niveaux d'analyse tels que définis classiquement (notamment les niveaux phonologique, morphologique, syntaxique et sémantique) :

- des *propriétés combinatoires*, qui concernent la façon dont ces unités élémentaires s'intègrent à ces énoncés,
- des *propriétés relationnelles*, qui identifient des régularités dans les relations entre unités élémentaires.
- des *propriétés intrinsèques*, propriétés qui sont indépendantes des autres éléments constituant les énoncés dans lesquels elles interviennent,

Il n'y a pas de raison *a priori* de penser qu'une même segmentation produise des unités élémentaires qui soient toutes les récipiendaires légitimes de tous les types de

— Langacker (1972, p. 37) : « Le mot est une notion difficile à définir » (*The word is a difficult notion to define*) ;

— Matthews (1991, p. 208) : « De nombreuses définitions du mot ont été proposées, et si l'une d'entre elles avait été un succès, je l'aurais donnée depuis longtemps, plutôt que d'esquiver la question jusqu'ici » (*There have been many definitions of the word, and if any had been successful I would have given it long ago, instead of dodging the issue until now*).

2. Nous écartons par ailleurs les langues des signes de notre discussion, certaines de leurs propriétés fondamentales les différenciant des autres langues, singulièrement sur les questions abordées ici.

propriétés que l'on peut être amené à identifier et à décrire. Cette approche nous permettra de proposer plusieurs définitions distinctes correspondant ainsi à différents types de « mots ». Chacun des travaux décrits dans le reste de ce document s'articule autour de l'un ou l'autre de ces types de mots, parfois autour de l'articulation non triviale entre deux types de mots différents. Toutefois, les réflexions proposées dans ce premier chapitre sont très limitées en regard de l'ampleur des problèmes évoqués. Si elles sont un préalable nécessaire à tous les chapitres qui suivent, elles ne font qu'effleurer de nombreuses questions qui dépassent largement le cadre de notre travail.

Naturellement, plusieurs unités élémentaires peuvent partager certaines propriétés. Un *lexique* est alors un inventaire formé de tels ensembles d'unités élémentaires, appelées *unités lexicales*, chaque unité lexicale étant associé à des propriétés communes à toutes les unités élémentaires qui la composent (dans le cas de propriétés intrinsèques ou combinatoires) ou qualifiant les relations entre ces unités élémentaires (dans le cas de propriétés relationnelles). Chaque unité lexicale ainsi munie de propriétés linguistiques dans un lexique est appelée *entrée lexicale*. Les unités lexicales présentes dans un lexique sont souvent identifiées, au moins en partie, par l'une des unités élémentaires qui les composent : c'est ce que l'on appelle la *forme de citation* de l'unité lexicale.

Les unités lexicales doivent être *minimales* pour chacune des propriétés qui leur sont attribuées, les propriétés des combinaisons d'unités élémentaires au sein des énoncés étant décrites non pas dans un lexique mais dans une *grammaire* de la langue. *A contrario*, le fait qu'une unité lexicale puisse porter par elle-même des propriétés, propriétés qui sont indépendantes du contexte d'utilisation de l'unité lexicale, lui confère ce que nous appelerons son *autonomie*.

Un lexique regroupe généralement des unités lexicales associées à des propriétés de nature identique. Anticipant sur la suite de ce chapitre, on peut par exemple constituer un lexique rassemblant des unités lexicales qui, pour simplifier, sont chacune associées à un sens intrinsèque. Ces entrées lexicales forment alors un lexique qualifé de *lexique sémantique*. Par exemple on peut regrouper dans une même unité lexicale des unités élémentaires du français telles que *voler*, *volent* et *voleront* issues de la segmentation d'énoncés au sein duquel elles ont un sens qui correspond à celui de l'anglais *to fly*. On peut par convention nommer VOLER$_1$ cette unité lexicale (« voler » est donc sa forme de citation) et lui associer une représentation du sens correspondant, constituant ainsi une entrée lexicale sémantique. Les unités élementaires *voler*, *volent* et *voleront* issues de la segmentation d'énoncés au sein duquel elles ont un sens qui correspond à celui de l'anglais *to steal* peuvent de même être regroupées en une unité lexicale VOLER$_2$ elle aussi associée à une représentation de son sens. Si en revanche on s'intéresse aux propriétés flexionnelles, propriétés relationnelles sur lesquelles nous reviendrons également par la suite, l'ensemble de ces unités élementaires, indépendamment de leur sens, peuvent

être regroupées en une unité lexicale unique au sein de laquelle elles sont reliées par un ensemble unique de propriétés relationnelles que l'on retrouve pour d'autres verbes dits « du premier groupe », puisque, pour simplifier, VOLER$_1$ et VOLER$_2$ se conjuguent de la même façon. L'entrée lexicale unique ainsi obtenue a alors sa place dans un *lexique flexionnel*.

Cette définition de la notion de lexique nous permet d'utiliser le terme générique de *propriété lexicale* pour dénoter une propriété intrinsèque, combinatoire ou relationnelle. Naturellement, une même séquence de phones ou de graphèmes peut faire partie, dans un lexique donné, de plusieurs unités lexicales distinctes associées à des propriétés distinctes. Dans le cas d'un lexique sémantique, c'est par exemple le cas des homonymes.

Idéalement, et pour une langue donnée, on pourrait souhaiter que toutes les propriétés lexicales définies par un modèle formel de cette langue puissent être associées à des unités lexicales pour lesquelles elles seraient toutes simultanément pertinentes, malgré la diversité voire la disparité de ce qu'elles décrivent. Autrement dit, on pourrait segmenter les énoncés d'une langue en unités élémentaires d'une façon unique, les unités ainsi obtenues étant toutes les récipiendaires légitimes de toutes les propriétés lexicales identifiées, quelle que soit leur nature (prosodique, morphologique, syntaxique, sémantique, etc.). C'est du reste ce que semblent impliquer Booij et Hulk (1988) lorsque, décrivant la distinction entre grammaire et lexique dans (Chomsky, 1965), ils définissent le lexique comme « un ensemble non ordonné (une liste) d'entrées lexicales vues comme des ensembles de traits : traits phonologiques, sémantiques et syntaxiques ». Une telle hypothèse est également au fondement de la notion de « mot » qui fait partie du langage commun [3, 4].

En réalité, comme nous allons le voir, cette hypothèse ne tient pas. C'est pour cette raison qu'il est notoirement difficile de proposer une définition linguistiquement pertinente, opératoire et générale de la notion de « mot », terme pourtant largement utilisé dans la littérature avec un sens qui se rapproche tantôt de la notion d'unité élémentaire, et

3. Ainsi, le Trésor de la Langue Française identifie pour le premier sens de *mot* deux définitions complémentaires, l'une présentée comme générale, l'autre comme relevant de la linguistique en tant que domaine de spécialité. La première définition fait du mot un « son ou groupe de sons articulés ou figurés graphiquement, constituant une unité porteuse de signification à laquelle est liée, dans une langue donnée, une représentation d'un être, d'un objet, d'un concept, etc. » La superposition entre unité articulatoire, unité graphique et unité sémantique est donc explicite. La définition du sens linguistique du mot ne reproduit pas cette superposition, mais est le lieu d'une certaine démission : « unité significative indépendante, ne pouvant pas toujours être déterminée selon un critère de séparabilité fonctionnelle ni par un critère de délimitation intonative. » La définition est donc purement sémantique — bien que la notion d'indépendance sémantique n'y soit pas définie —, l'applicabilité de critères non sémantiques étant indiquée comme n'étant pas systématique. Mais aussitôt, le Trésor de la Langue Française illustre cette définition par une citation de Saussure dans laquelle *porte-plume* est explicitement considéré comme étant « une unité plus large qu'[un mot] », en dépit de son indépendance sémantique manifeste.

4. On notera que l'étymologie de *mot* reflète ironiquement son peu de pertinence : *mot* est dérivé du latin populaire **mottum*, altération de *muttum*, attesté dans l'expression *nullum mutum* 'pas un mot', littéralement 'pas même un grommèlement (*mu*)' (Bloch et von Wartburg, 1932). Cf. aussi *muttīre* 'grommeler', dérivé lui aussi de l'onomatopée *mu*.

tantôt de la notion d'unité lexicale, souvent en ignorant le fait que différentes propriétés lexicales peuvent n'être pas toutes pertinentes simultanément sur une unité donnée. La difficulté qui en résulte à proposer une définition de la notion de mot est discutée tant dans les manuels que dans les publications de résultats de recherche, et ce dans de nombreux domaines tels que la morphologie, la syntaxe ou la typologie. On peut à cet égard se reporter à Tesnière (1959, ch. 10), Matthews (1974, p. 208), Tournier (1988, p. 9), Joseph (2001), Aronoff et Fudeman (2005, p. 34), Haspelmath (2011), ou au volume complet consacré à une discussion typologique de la notion de mot, édité par Dixon et Aikhenvald (2003). Si donc il est difficile voire impossible de définir ce qu'est un mot, est-il possible de définir de façon cohérente une notion d'unité lexicale, et si oui, de quelle façon ? En réalité, il nous semble que seule une analyse des propriétés lexicales elles-mêmes peut permettre de cerner par une approche ascendante les unités auxquelles elles s'appliquent.

Plutôt que nous attarder dans l'étude de multiples définitions de la notion de mot proposées par différents auteurs [5], que ce soit pour une langue donnée ou dans une perspective typologique, la direction que nous allons suivre va donc consister à prendre pour point de départ les propriétés lexicales elles-mêmes. Les unités linguistiques récipiendaires de chacune de ces propriétés lexicales formeront ainsi les instances d'*une* notion d'unité lexicale parmi plusieurs. Cette approche permettra d'identifier les différents types d'unités lexicales sur lesquelles nous avons travaillé et sur lesquels s'appuient les différents chapitres formant la suite de ce document, en gardant en tête le cadre qui est le nôtre, celui du traitement automatique des langues. Naturellement, les définitions que nous serons amenés à proposer seront des abstractions idéalisées. Il est peut-être possible de proposer, pour certaines d'entre elles, des formulations plus précises qui seraient applicables de façon rigoureuse et indépendante de la langue, mais cela dépasse largement le cadre de ce document.

Cette approche nous conduit inévitablement à nous écarter de l'intuition que l'on peut avoir sur ce que peut être un mot dans une langue que l'on maîtrise. Avant de discuter des propriétés lexicales et de la façon dont leur étude peut permettre de proposer des définitions pour plusieurs types d'unités lexicales, il est donc utile de discuter rapidement de ce que pourrait recouvrir la notion *intuitive* du mot.

5. Et ce souvent en tombant dans une forme rédhibitoire de circularité, les définitions de la notion de mot reposant sur des critères qui ont tendance à s'appuyer sur une notion de mot préalable. Par exemple, il est inapproprié de définir un mot comme étant un représentant d'une classe distributionnelle cohérente si l'on définit les classes distributionnelles à partir d'une segmentation préalable en mots — cf. par exemple la critique de Kahane (2008) par Magistry (2013, p. 36). Ce n'est cependant pas la seule façon de procéder pour faire s'appuyer la notion de mot sur celle de classe distributionnelle.

1.2 Le « mot », une notion intuitive ?

Il est une idée étrangement répandue : celle selon laquelle il existerait une notion intuitive du mot. Kočourek (2001, p. 5) et Haspelmath (2011, p. 3) citent ainsi Coseriu (1964) : « Nous estimons la notion de 'mot' comme intuitivement établie. » Sapir (1921) ne disait pas autre chose quand il écrivait qu'il « n'y a pas la moindre difficulté à prendre conscience du mot en tant que réalité psychologique » [6]. Cette intuition de ce qu'est un mot peut nous sembler en effet naturelle, quand bien même elle contraste avec les difficultés à formuler une définition précise. C'est ce qu'écrivent Aronoff et Fudeman (2005, p. 34) : « […] ce qui semble être une tâche relativement simple (nous pensons tous savoir ce qu'est un mot, non ?) s'avère si problématique » [7]. De même Sapir (1921) : « il est plus facile de poser la question [de ce qu'est un mot] que d'y répondre » [8]. Il y a au moins trois directions dans lesquelles on peut chercher des éléments de réponse face à ce paradoxe apparent.

La première vient de la linguistique de terrain. En effet, Sapir (1921) indique que des « [amér]indiens naïfs » [9], non lettrés, peuvent de façon cohérente dicter un texte mot à mot, tout en refusant de discuter d'unités plus petites que le mot, au prétexte qu'elles n'ont aucun sens. Ce dernier point contraste avec le constat inverse de Evans *et al.* (2008), qui ont constaté sur le dalabon (gunwinyguan, macro-gunwinyguan, Australie), une langue polysynthétique, que des pauses à l'intérieur de ce qu'ils analysent comme étant des mots sont possibles, quoique plus courtes qu'entre les mots et placées à des positions contraintes à la fois morphologiquement et phonologiquement. On peut également citer (Bloomfield, 1933a, p. 178) : « les individus qui n'ont pas appris à lire et à écrire éprouvent des difficultés lorsque, par hasard, il leur est demandé d'indiquer des frontières entre mots » [10]. On pourra également noter que l'histoire des langues écrites est pleine d'évolutions graphiques qui reflètent ces difficultés (cf. l'anglais *another*, qui correspond sémantiquement à *an other*, orthographe aujourd'hui incorrecte [11]).

6. *[…] there is not […] the slightest difficulty in bringing the word to consciousness as a psychological reality.*

7. *[…] what seems like a relatively simple task (we all think we know what a word is, right ?) proves to be so problematic.*

8. *It is easier to ask the question than to answer it.*

9. *[…] the naïve Indian […]*

10. *People who have not learned to read and write, have some difficulty when, by any chance, they are called upon to make word divisions.*

11. La fusion est aujourd'hui si installée qu'on entend parfois *a whole 'nother* 'un tout autre' (au lieu de *a whole other*), résultat d'une fausse coupe (ici, une tmèse). Plus simplement, de telles fausses coupes conduisent souvent, dans l'histoire des langues, à la création de nouveaux mots étymologiquement inattendus. En voici quelques exemples. L'anglais *nickname* 'surnom' < moyen-anglais *nekename* résulte de la fausse coupe (ou métanalyse) de *an ekename* 'un surnom' (cf. aussi l'anglais *newt* 'triton' issu de *an ewt(e)*, ou le français *lierre*, résultat de l'agglutination de l'article défini à un ancien *hyerre* ∼*ierre*). L'inverse s'est produit dans le cas de l'anglais *apron* 'napperon', emprunt à l'ancien français *napperon* avec réinterprétation de *a napron* en *an apron* (cf. aussi le mot français *mie*, dans *ma mie*, métanalyse de *m'amie*, où le possessif *ma* est élidé). L'un des exemples les plus amusants de ce type de phénomène est le cas du *hamburger*, étymologiquement dérivé du nom de la ville de Hambourg, mais métanalysé en *ham+burger*, d'où le mot *burger* et les composés secondaires de type *cheeseburger*. Et pourtant, il n'y a pas de jambon (anglais *ham*) dans un hamburger.

La seconde manière dont on a pu tester l'existence d'une notion intuitive du mot vient également de la linguistique de terrain. Il s'agit de l'idée consistant à étudier l'intuition que peuvent avoir des locuteurs d'une langue non écrite mais qui maîtrisent une autre langue et son système d'écriture. C'est ce qu'a réalisé Peterson (2008), cité par Haspelmath (2011), sur le kharia (munda, austroasiatique, Inde). Six locuteurs de cette langue qui maîtrisent le hindi écrit ont eu à identifier les mots composants 12 énoncés. Ces locuteurs ont fourni des résultats très différents, allant jusqu'à identifier entre un et quatre mots pour une même phrase. Malgré le biais favorable qu'aurait pu constituer la maîtrise du hindi écrit, la notion intuitive de mot n'était donc pas cohérente d'un locuteur à l'autre.

Un troisième angle, *a contrario*, est fourni par le constat de l'influence des systèmes d'écriture, lorsqu'ils existent, sur la façon dont se construit individuellement et sociologiquement une notion de mot. Dans les langues occidentales modernes, l'écriture se fait au moyen de l'alphabet latin, et compter les mots dans un texte consiste souvent à compter, comme ci-dessus, le nombre de séquences de caractères allant d'un séparateur, l'espace, à un autre. Nous reviendrons sur cette dimension plus en détail à la section 7. Mais est-il raisonnable de penser que des conventions typographiques héritées du passé soient en accord avec une segmentation linguistiquement pertinente en synchronie ? En français, l'exemple de *au fur et à mesure* montre que la réponse est négative, *fur* n'ayant en synchronie aucune existence propre. Un autre cas intéressant est celui du mandarin (sinitique, sino-tibétain, Chine), qui utilise un système d'écriture faisant usage de caractères idéographiques (sinogrammes) mais pas d'un séparateur d'unités graphiques comparable à l'espace en français (nous y reviendrons là aussi à la section 7). Comme le montre notamment DeFrancis (1984), le statut culturel du sinogramme a conduit pendant longtemps à identifier les notions de sinogramme et de mot, mais également de syllabe, tous les sinogrammes correspondant à une monosyllabe [12]. Cette position, héritée du chinois classique, n'est plus valable aujourd'hui. DeFrancis (1984) propose ainsi de distinguer les sinogrammes libres, semi-liés et liés : un mot, quelle que soit l'acception que l'on donne à ce terme, peut alors être composé de plusieurs sinogrammes. La tâche consistant à identifier ces unités dans des textes se présente alors. Si l'on se met d'accord sur une série de règles et de critères définissant le mot d'une façon reproductible et cohérente, on peut alors, manuellement ou automatiquement, réaliser cette tâche de segmentation — on notera qu'au moins quatre corpus ainsi segmentés manuellement existent aujourd'hui pour le mandarin, et suivent des conventions de segmentation incompatibles les unes par rapport aux autres, ce qui contribue une fois

12. Cette approximation est presque parfaitement exacte. Les très rares contre-exemples sont (i) le sinogramme 兒 / 儿 , qui dans certains cas n'a comme seule valeur phonologique que d'indiquer une rétroflexion de la syllabe précédente et (ii) certains sinogrammes tombés en désuétude créés pour dénoter des unités de mesure occidentales.

encore à mettre en doute l'existence d'une notion intuitive du mot [13, 14]. On peut aussi proposer des modèles mathématiques ou informatiques de ce que pourrait être un mot, l'appliquer sur des textes et étudier la nature et la pertinence des unités ainsi extraites. Nous y reviendrons à la section 7.4. Des modèles distincts pouvant conduire à plusieurs segmentations différentes, le fait que nous puissions être amenés à définir plusieurs notions d'unités lexicales prend ici un écho intéressant.

Il s'avère donc impossible de ressortir à l'intuition des locuteurs pour tenter de délimiter une notion opératoire du mot [15, 16]. Comme annoncé précédemment, nous devons donc nous en remettre à l'étude des différents types de propriétés lexicales que nous pourrions vouloir attribuer à une telle unité, en prenant en compte à la fois la *minimalité* et l'*autonomie* de ces dernières — la minimalité et l'autonomie se déclinant de façon différente d'une propriété lexicale à l'autre.

1.3 Propriétés lexicales

Avant de passer en revue un certain nombre de propriétés qu'il est loisible de chercher à attribuer à une unité élémentaire ou à une unité lexicale, précisions un point. Contrairement à ce qui se passerait si nous cherchions à définir directement une notion d'unité élémentaire ou d'unité lexicale, il est possible de faire reposer les propriétés lexicales sur une segmentation préalable en unités sans pour autant introduire de circularité dans les définitions. En effet, on peut envisager qu'un énoncé soit segmenté en unités de multiples manières et que soit étudiée l'applicabilité de chaque propriété lexicale envisagée à l'ensemble des unités produites par chacune de ces segmentations. Une segmentation cohérente est alors une segmentation telle que toutes les unités qu'elle induit sont des récipiendaires légitimes d'une propriété lexicale particulière, laquelle propriété lexicale peut parfaitement être définie pour l'une de ces unités en fonction des autres unités produites par la même segmentation de l'énoncé.

13. On peut également mentionner le pinyin, système de romanisation du mandarin adopté par l'Organisation internationale de normalisation (ISO), qui utilise l'espace comme séparateur de « mots ». Il a donc été nécessaire de développer des règles définissant ce qu'est un « mot » (norme chinoise GB/T 16159-2012). Ces règles combinent des critères sémantiques, formels (règles régissant la façon dont sont écrits les différents types de mots à réduplication), les préfixes et suffixes, morphosyntaxiques (selon la partie du discours), etc.

14. Du reste, certains auteurs mettent en doute la pertinence de la notion même de mot, du moins pour le mandarin (Liu *et al.*, 2013).

15. Nous ne voulons pas signifier par là que la notion de mot est dénuée de réalité psycholinguistique, mais simplement que cette réalité, si elle existe, n'est pas intuitivement accessible au locuteur. En réalité, il a été montré par exemple que des informations fréquencielles sur les mots (avec différentes définitions de ce qu'est un mot) influencent les temps de réponse pour différentes tâches (Baayen, 2009). Il semble donc exister un type d'unité dont les propriétés, notamment fréquencielles et distributionnelles, influencent les locuteurs. De plus, ces derniers semblent également influencés par les relations qu'entretiennent ces unités avec d'autres qui lui sont reliées (Schreuder et Baayen, 1997 ; Milin *et al.*, 2009).

16. Pour une analyse plus détaillée, on pourra par exemple se référer à Blevins (2016, sec. 3.1.1).

Il en résulte que l'on peut définir autant de types d'unités lexicales que l'on a identifié de propriétés lexicales. C'est ce que nous ferons dans la suite de cette section. Nous n'avons pas pour ambition de passer en revue l'ensemble des propriétés lexicales, si tant est que cela soit possible. Pour un inventaire plus large et discuté de façon plus détaillé dans le contexte de l'étude de la notion de mot au niveau morphosyntaxique, on pourra ainsi se reporter à Haspelmath (2011).

Dans toute la suite de ce chapitre, et dans le reste de ce document, nous aurons recours à un abus de langage très répandu : nous utiliserons la même terminologie pour qualifier certains types d'unités élémentaires et les unités lexicales dans lesquelles elles sont incluses. Anticipant à nouveau sur la suite de ce chapitre, nous pourrons considérer qu'en français, les unités élémentaires *clémentine* et *clémentines* ont le même sens, du moins selon une propriété lexicale particulière, celle du sens intrinsèque. Nous utiliserons le qualificatif de *mot sémantique* pour de telles unités, et nous l'appliquerons d'une part aux unités élémentaires *clémentine* et *clémentines*, unités définies par rapport à cette propriété du sens intrinsèque, et d'autre part à l'unité lexicale qui rassemblerait donc ces deux unités élémentaires dans un lexique sémantique reposant sur cette même propriété. Cet abus de langage se justifie par le fait que cette unité lexicale rassemble justement des unités élémentaires ayant le même sens intrinsèque.

Dans la suite de cette section, nous allons donc survoler différents types de propriétés lexicales. L'identification des différents types de propriétés lexicales et les termes que nous emploierons pour les nommer ne reprennent pas telle quelle une théorie établie mais s'appuient néanmoins souvent sur de nombreux travaux antérieurs, dont l'étude, la comparaison et la critique approfondies dépasserait de loin le cadre de ce document. Notre objectif n'est ici que d'aboutir à des ébauches de définitions pour différents types d'unités lexicales que nous aurons à manipuler dans la suite du document.

1.3.1 Sens intrinsèque

Traditionnellement, la propriété la plus naturelle à associer à une unité lexicale est un sens. Notons que nous ne nous préoccupons pas ici de savoir comment ce sens peut-être représenté [17], mais seulement au fait que l'on veuille associer un sens à des unités. Dans un premier temps, le sens auquel nous ferons référence est le *sens intrinsèque*, pour ainsi dire dictionnairique, c'est-à-dire indépendant d'autres unités environnantes — les propriétés sémantiques combinatoires font l'objet d'une section ultérieure. Combiné avec le critère de minimalité, défini plus haut, on obtient une propriété qui correspond à des formulations classiques du mot telles que les suivantes :

17. Le panel des représentations possibles est très large, et peut aller entre autres d'une représentation vectorielle dans une espace continu (Turian *et al.*, 2010 ; Collobert *et al.*, 2011 ; Mikolov *et al.*, 2013 ; Vulić et Korhonen, 2016) à des formules logiques impliquant des quantificateurs du second ordre, par exemple pour les déterminants du français.

> […] le mot est l'un des éléments de « sens » individuels minimaux et complètement cohérents dans lesquels la phrase se décompose [18]. (Sapir, 1921, p. 34)

> [L'unité est] une tranche de sonorités qui est, à l'exclusion de ce qui précède et de ce qui suit, le signifiant d'un certain concept. (de Saussure, 1916, p. 146)

On peut noter que la minimalité peut être reformulée dans ce cas en termes de compositionnalité sémantique : identifier une unité comme récipiendaire pertinent d'un sens implique qu'il n'en existe pas de décomposition en sous-unités telles que *toutes* ces sous-unités puissent se voir attribuer un sens dont la composition régulière produise le sens de l'unité dans son ensemble. C'est ce que Gross (1996) appelle le caractère *opaque* d'une telle unité, et définit les *decoding idioms* de Makkai (1972). L'autonomie, quant à elle, implique que toute unité à laquelle est attribué un sens puisse être combinée, selon des règles grammaticales, avec d'autres unités, le résultat ayant pour sens le résultat de la composition des sens de l'ensemble de ces unités.

Amalgamée à une propriété orthogonale, la possibilité de former un énoncé complet, la notion de sens intrinsèque est également sous-jacente à la définition de Bloomfield (1933b, p. 156) :

> Une forme libre minimale est un *mot*. Un mot est ainsi une forme qui peut être énoncée en isolation (tout en faisant sens) mais ne peut être analysée en parties qui (toutes) peuvent être énoncées en isolation (tout en faisant sens) [19].

Par définition, le sens intrinsèque d'une unité élémentaire n'est pas altéré par les propriétés qui régissent la façon dont cette unité se combine à d'autres pour former un énoncé, qu'il s'agisse de propriétés décrivant la façon dont l'unité s'insère sémantiquement dans son environnement ou la façon dont les propriétés sémantiques de l'unité imposent des contraintes à d'autres unités de son environnement — nous reviendrons ci-dessous sur ces autres propriétés sémantiques. Ainsi, pour reprendre l'exemple ci-dessus, les unités élémentaires *clémentine* et *clémentines* issues respectivement de la segmentation des énoncés *je mange une clémentine* et *je mange des clémentines* partagent un même sens intrinsèque, mais diffèrent par la façon dont ce sens intrinsèque s'insère dans le sens global de l'énoncé (singulier vs. pluriel). Il en va de même des unités élémentaires slovaques *kartu* et *kartou* dans les énoncés *mám kartu* 'j'ai une carte (de crédit)' et *platím kartou* 'je paye par carte'. Sans surprise, nous utiliserons les termes de *mot sémantique* ou de *lexème* pour dénoter une unité lexicale regroupant des unités élémentaires de même sens intrinsèque. Suivant également les conventions

18. *[…] word is one of the smallest, completely satisfying bits of isolated "meaning" into which the sentence resolves itself.*

19. *A minimum free form is a* word. *A word is thus a form which may be uttered alone (with meaning) but cannot be analyzed into parts that may (all of them) be uttered alone (with meaning).*

habituelles, nous dénoterons un mot sémantique par l'une des unités élémentaires qui le composent (choisie conventionnellement) au moyen d'une typographique dédiée, pour le distinguer desdites unités élémentaires. Ainsi, nous écrirons que les unités élémentaires ci-dessus *clémentine* et *clémentines* font partie du mot sémantique ou du lexème CLÉMENTINE.

La difficulté principale avec cette propriété lexicale réside dans ce qu'il n'y a de moyen clair ni pour savoir ce qui porte un sens et ce qui n'en porte pas (cf. la préposition *de* en français [20]), ni pour distinguer les sens entre eux [21], ni pour expliciter ce qu'est une composition régulière.

Il n'en reste pas moins, pour prendre un exemple clair en français, que l'on peut utiliser l'unité *table ronde* pour dénoter une sorte de réunion-débat qui n'est ni une *table* ni *ronde*. À cet égard, *table ronde*, avec ce sens, est une unité à laquelle il est pertinent d'associer un sens, qui correspond donc à un mot sémantique [22].

Considérons désormais l'énoncé *Pierre a mis de l'eau dans son vin*. Si l'on essaye de le décomposer en unités minimales et autonomes légitimes à recevoir un sens, on s'aperçoit que l'ensemble *mis de l'eau dans son vin* n'est pas décomposable. L'ensemble *mis de l'eau dans son vin* correspond donc un mot sémantique, qui n'est pas nécessairement toujours d'un seul tenant (*Pierre a mis à nouveau de l'eau dans son vin*). Autrement dit, les unités lexicales peuvent être discontinues — c'est du moins le cas des mots sémantiques.

1.3.2 Sens conventionnel

Une autre propriété lexicale que l'on peut vouloir attribuer à une unité est son *sens conventionnel*, c'est-à-dire l'information pragmatique selon laquelle il est convenu entre les locuteurs d'une même langue, souvent entre ceux d'entre eux qui sont les spécialistes d'une discipline particulière, qu'un certain sens (ou concept) est dénoté par cette unité. La différence avec le sens intrinsèque se situe au niveau de la notion de minimalité. En effet, il se peut qu'une unité minimale au titre du sens conventionnel ne le soit pas au titre du sens intrinsèque, en tant qu'elle est décomposable en sous-unités dont les sens intrinsèques respectifs se combinent compositionnellement pour former le sens intrinsèque global de l'unité complète. Cette situation est celle des *encoding idioms* de Makkai (1972).

20. Par exemple, dans *je viens de Bourgogne*, le *de* porte un sens en relation avec la notion de provenance ; mais dans *j'essaye de dormir*, le *de* n'a pas à proprement parler de sens.

21. Nous verrons par exemple au chapitre 6 quel choix a été fait en ce sens par les développeurs de lexiques sémantiques de type wordnet, et notamment le Princeton WordNet (Fellbaum, 1998).

22. Il en est de même, pour reprendre des exemples de Kahane (2008), d'un mot comme *aspirateur*, malgré sa décomposabilité en *aspir-* et *-ateur*. En effet, quand bien même il serait possible d'attribuer un sens à l'élément -ATEUR (X-ATEUR signifiant, en tout cas ici, 'appareil servant à X-er'), sa combinaison avec un élément comme *aspir-* n'est pas libre : *LAVATEUR, *NETTOYATEUR. Autrement dit, *-ateur* n'est pas autonome, et ce n'est qu'à l'unité *aspirateur* qu'il convient d'attribuer un sens intrinsèque. Autrement dit, *aspirateur* est un mot sémantique, mais *-ateur* n'en est pas un.

Étudions ceci sur un exemple. Soit l'énoncé *Pierre s'est acheté une nouvelle machine à laver*. L'énoncé *Pierre s'est acheté un nouvel appareil à laver* (voire *appareil à nettoyer*) est *a minima* peu compréhensible et peu naturel, bien qu'*appareil* et *machine* aient, dans ces deux énoncés, des sens intrinsèques très proches voire identiques. C'est parce que le découpage pertinent par rapport à la propriété du sens conventionnel fait de *machine à laver* une unité lexicale, qui respecte, pour cette propriété particulière, les contraintes d'autonomie et de minimalité [23]. Nous appellerons *termes* de telles unités. Mais un sens intrinsèque peut être attribué aux trois unités *machine*, *à* et *laver*. La spécification de la notion de minimalité dans le cas du sens conventionnel est en général dénotée par le terme de *unithood* dans la littérature sur l'extraction automatique de terminologies (Kageura et Umino, 1996).

Le sens conventionnel est la propriété lexicale au cœur du modèle wordnet, dont le premier exemple est le Princeton WordNet (PWN ; Fellbaum, 1998). Nous reviendrons sur ce modèle au chapitre 6 lorsque nous décrirons nos travaux sur le développement de wordnets pour le français et le slovène. Dans les travaux sur les wordnets, il est d'usage d'utiliser le terme de *littéral* (anglais 'literal') pour dénoter une unité lexicale à laquelle on cherche à attribuer un sens intrinsèque. Dans d'autres contextes nous qualifierons plutôt de *terme* une unité lexicale (ou élémentaire) définie comme un récipiendaire légitime d'un sens conventionnel. On notera que la majorité des termes sont également des mots sémantiques (au sens du paragraphe précédent). Traditionnellement, on n'appelle souvent « terme » que les termes qui, comme *machine à laver*, ne sont *pas* également des mots sémantiques. Nous nous rangerons à cet usage et utiliserons donc le terme de *mot sémantique* pour dénoter toutes les unités dont le sens intrinsèque est autonome, quand bien même il n'est pas minimal. Autrement dit, et par abus de langage, nous qualifierons de *mot sémantique* à la fois un mot sémantique (au sens du paragraphe précédent) comme MACHINE$_{\text{E}}$ t un terme (*stricto sensu*) comme *machine à laver*.

1.3.3 Sens combinatoire

Un autre volet du sens que l'on peut vouloir associer à une unité lexicale concerne son *sens combinatoire*, c'est-à-dire les spécificités de cette unité quant à la façon dont elle se combine à d'autres pour construire un sens compositionnel. Comme nous l'avons déjà indiqué, le sens combinatoire ainsi défini peut se décomposer en deux parties : le *sens combinatoire interne*, qui rassemble les propriétés qui régissent la façon dont cette unité s'insère dans son environnement, et le *sens combinatoire externe*, qui décrit la façon dont l'unité impose des contraintes à d'autres unités de son environnement.

23. On notera que dans l'expression *faire une machine*, le segment *à laver* peut être élidé, à condition que le contexte le permette.

1.3.3.1 Actance

Pour les unités dont le sens intrinsèque exprime un procès, la *structure actancielle*, ou *structure prédicative*), modélise l'inventaire des participants au procès qu'elle dénote. Pour d'autres types d'unités, la structure actantielle (terme utilisé alors avec un sens étendu) peut être d'une nature légèrement différente, notamment pour les unités porteuses d'informations de quantification.

Restons-en aux unités prédicatives. Par exemple, on peut vouloir décrire le fait qu'un verbe anglais comme OFFER 'offrir', ou du moins son sens le plus fréquent, implique trois participants, ou *actants*, à savoir la personne qui offre quelque chose, la personne à qui elle est offerte et la chose offerte. Il s'agit de propriétés sémantiques constitutives de ce qu'est l'acte d'offrir, et qui ont un impact sur la façon dont OFFER se combine avec d'autres unités pour former un sens cohérent à l'échelle d'un énoncé. C'est là le type de propriétés lexicales qu'encode une ressource comme FrameNet (Baker *et al.*, 1998).

Considérons désormais le verbe *casser* dans son acception la plus fréquente de 'mettre en morceaux'. En français, il peut prendre part à plusieurs types d'énoncés, dont deux sont illustrés respectivement par *Pierre a cassé le vase* (emploi transitif) et par *Le vase a cassé* (emploi neutre). Nous retombons naturellement dans la difficulté, mentionnée plus haut, qu'il y a à décider si deux sens sont identiques ou distincts. Il semble néamnoins raisonnable de choisir de faire de ces deux occurrences de la chaîne graphémique *cassé* deux unités ayant le même sens intrinsèque : ce sont deux occurrences du même mot sémantique. Pourtant, elles ne partagent pas le même sens combinatoire, l'une ayant deux arguments et l'autre un seul. Nous dirons qu'il s'agit de deux *mots actanciels* distincts.

Soit un autre exemple, lui aussi classique : *Mon libraire vend ce livre sans problème* (emploi transitif) en parallèle avec *Ce livre se vend sans problème* (emploi *se*-moyen). Comme nous le verrons à la section 5.1.3, il est possible de voir dans l'emploi *se*-moyen une construction ne permettant pas d'exprimer l'agent du procès mais où l'agent continue à faire partie des actants exprimés. Sous cette interprétation, les deux occurrences de *vend* ont à la fois le même sens intrinsèque et le même sens combinatoire, contrastant ainsi avec l'exemple précédent : nous sommes en présence de deux occurrences du même mot sémantique et du même mot actanciel.

De telles alternances, ainsi que d'autres types d'alternances parfois plus complexes, font l'objet d'innombrables travaux — cf. en particulier ceux de Levin (1993), Gross (1975) ou Boons *et al.* (1976a,b). Il en a résulté certaines ressources lexicales organisées autour de ces alternances, comme VerbNet pour l'anglais (Kipper Schuler, 2005), mais également des ressources telles que les tables du Lexique-Grammaire (Gross, 1975 ; Boons *et al.*, 1976a,b).

1.3.3.2 Traits sémantiques

Comme évoqué précédemment, plusieurs unités élémentaires peuvent partager un même sens intrinsèque. Ces unités élémentaires peuvent différer par leur structure actancielle, comme nous venons de le voir, mais également par des *traits sémantiques* tels que le nombre (ainsi *clémentines* vs. *clémentine*), le cas, l'aspect, etc. Il faut bien distinguer ces traits des traits morphologiques que nous évoquerons plus bas, bien que les termes employés (ainsi le terme de « nombre » ou de « pluriel ») soient souvent les mêmes (cf. le phénomène de *décalage*, sur lequel nous reviendrons au chapitre 2, illustré par les verbes dits « déponents » du latin, morphologiquement passifs mais sémantiquement actifs). Les phénomènes d'accord, quant à eux, peuvent mettre en jeu à la fois les traits sémantiques et les traits morphologiques, ainsi que d'autres propriétés (Corbett, 2006).

Pour reprendre l'un des exemples évoqués ci-dessus, les unités élémentaires slovaques *kartu* et *kartou* dans les énoncés *mám kartu* 'j'ai une carte (de crédit)' et *platím kartou* 'je paye par carte' ont un même sens intrinsèque, celui de carte de crédit, mais s'insèrent selon des modalités différentes au sein de la structure sémantique de l'énoncé.

Par définition, les traits sémantiques ont donc pour récipiendaire légitime une unité dotée d'un sens intrinsèque, c'est-à-dire un mot sémantique.

1.3.4 Comportement positionnel et combinatoire dans l'énoncé

Une langue peut être considérée comme un système permettant de coder un sens (éventuellement complexe) sous formes d'énoncés linguistiques (au sens défini au début de ce chapitre). La production d'un énoncé est ainsi une opération d'encodage et son interprétation une opération de décodage — chacune de ces opérations étant inévitablement imparfaite. L'un des points communs entre toutes les langues est que les systèmes de codages qu'elles constituent produisent tous des messages structuré. La séquentialité des productions langagières induit ainsi une structuration *temporelle*. Mais le sens qu'exprime un énoncé peut également être analysé comme étant structuré de façon *a minima* hiérarchique, induisant la possibilité d'une structuration *hiérarchique* de l'énoncé lui-même.

À l'échelle d'un énoncé (voire d'une séquence d'énoncés), la structuration temporelle peut être analysée par la définition de *constituants*, des séquences contiguës d'unités élémentaires. La structuration hiérarchique, quant à elle, peut être analysée par la définion de relations de *dépendances* entre unités élémentaires. Il serait difficile, et en tout cas bien trop ambitieux ici, de proposer une définition universelle de ce que sont les constituants et les dépendances. Nous noterons simplement que la notion de constituant gagne à être définie comme une séquence *contiguë* de segments, l'utilisation de la notion de constituant discontinu relevant à notre sens d'une volonté d'y faire rentrer des considérations relevant en réalité de la notion de dépendance.

L'analyse de ces deux modes de structuration des énoncés (ou séquences d'énoncés) permet d'identifier des généralisations spécifiques à chaque langue. L'étude de ces généralisations constitue la discipline syntaxique. Il est désormais classique de décrire ces généralisations de façon formelle (règles de réécriture, modèles probabilistes de placement relatif des segments), en faisant interagir d'une part des propriétés générales de la langue et des propriétés particulières de certains segments. De telles descriptions s'appuient sur des segments élémentaires dont on analyse le comportement tant dans la structuration temporelle que dans la structuration hiérarchique. Il s'agit donc de deux types de propriétés qui, lorsqu'elles concernent des segments particuliers, sont des propriétés lexicales. Autrement dit, on peut considérer que les propriétés de structuration temporelle et les propriétés de structuration hiérarchique sont deux ensembles de propriétés, toutes deux syntaxiques, dont les récipiendaires légitimes sont des segments constituant deux types d'unités lexicales de niveau syntaxique.

1.3.4.1 Valence

Parmi les propriétés relevant de la structuration hiérarchique, la plus étudiée est la sous-catégorisation argumentale, ou *valence syntaxique*, qui permet un certain nombre de généralisations sur la façon dont les relations combinatoires entre un procès et ses actants, notions sémantiques relevant du mot actanciel défini plus haut, sont reflétées par la structuration hiérarchique au niveau syntaxique. Dans les cas les plus simples, ces relations correspondent directement à des dépendances de niveau syntaxique. Ce n'est pas toujours le cas. Il arrive ainsi que deux actants d'un même procès soient encodés non pas sous la forme de deux arguments, mais sous la forme d'un argument unique dont la structuration interne permet l'encodage des deux actants. C'est le cas d'un énoncé comme *Luc adore l'intelligence de Max*, où l'on peut identifier trois actants au procès principal : *Luc*, qui ressent l'adoration, *Max*, qui est le stimulus à l'origine de cette adoration, et *l'intelligence (celle de Max)*, une propriété de *Max* qui motive cette adoration (cf. *Luc adore Max pour son intelligence* ; Danlos *et al.*, 2016).

Nous utiliserons le terme de *mot syntaxique* pour désigner une unité qui correspond à la propriété de valence syntaxique. Il correspond à ce que certains auteurs appellent le *mot grammatical* (ainsi par exemple Blevins, 2016, sec. 3.2).

1.3.4.2 Catégorie morphosyntaxique

Parmi les propriétés relevant de la structuration temporelle, la plus classique est celle de *partie du discours*, ou *catégorie morphosyntaxique*. C'est souvent à partir des parties du discours que sont construites, grâce par exemple à la notion de constituance, des généralisations relatives à l'ordonnancement des segments dans l'énoncé. Là encore, les unités qui sont les récipiendaires légitimes d'une partie du discours ne correspondent

pas forcément à des unités qui sont les récipiendaires légitimes d'un sens intrinsèque, conventionnel ou combinatoire comme défini précédemment. Deux exemples en français :

— Certains des segments auxquels on attribue traditionnellement une partie du discours nommée « clitique » (ou une partie du discours plus spécifique au sein d'un ensemble de parties du discours regroupant plusieurs types de « clitiques ») peuvent être requis sans qu'il ne s'agisse de mots sémantiques, de mots actanciels ou de termes [24]. Ces unités respectent toutefois certaines généralisations relevant de la structure temporelle, notamment concernant leur placement par rapport aux verbes et aux autres unités. C'est par exemple le cas du *se* (ou *s'*) des verbes essentiellement pronominaux (s'ÉVANOUIR, SE BARRER), dont les règles de placement sont identiques à celles régissant les « clitiques » standard, c'est-à-dire ceux qui ont une contrepartie au niveau sémantique (placement de l'auxiliaire aux temps composés, placement relatif des différents « clitiques », etc.). C'est également le cas d'autres « clitiques », comme par exemple *en* (EN FINIR (AVEC)), *y* (S'Y CONNAÎTRE (EN)) ou *le* (ou *l'*) (L'EMPORTER) (van den Eynde et Mertens, 2006, p. 12).

— Un « mot composé » tel que *pomme de terre* a une structure syntaxique interne régulière, et on peut la découper en trois unités élémentaires à qui l'on peut attribuer une partie du discours. Pourtant, *pomme de terre* est un mot sémantique unique, son sens n'étant pas inférable à partir de sous-unités.

Nous parlerons de *mot morphosyntaxique* pour dénoter les unités (minimales) auxquelles on peut conférer une partie du discours. Les exemples ci-dessus montrent que certains types de mots composés sont des mots syntaxiques (et souvent sémantiques) mais sont composés de plusieurs mots morphosyntaxiques.

1.3.5 Propriétés flexionnelles

Les propriétés flexionnelles font l'objet de plusieurs chapitres de ce document (chapitres 2 à 4 et sections A.1 à A.4). Elles sont avant tout des propriétés relationnelles : elles identifient et qualifient des régularités dans les relations entre unités élémentaires. Il s'agit de régularités non-nécessairement systématiques qui concernent simultanément deux niveaux : le niveau formel et les traits sémantiques évoqués plus haut. Le caractère non-nécessaire de ces régularités implique que, par souci de cohérence du système de relations, on puisse considérer qu'un certain type de relation est instancié de façon valide (mais non-standard) même en l'absence de régularité formelle ou de régularité concernant les traits sémantiques. Considérons par exemple en français la relation entre *aimons* et *aiment*, entre *chantons* et *chantent* et entre *lisons* et *lisent*. La double régularité, au niveau de la forme et au niveau des traits sémantiques, est ici vérifiée. Il en va de même

24. On notera que le terme de « clitique » est mis ici entre guillemets. La raison en est que, dans le cas général, la notion de clitique est définie par rapport à des propriétés phonologiques (cf. ci-dessous).

quant aux relations entre *levons* et *lèvent* et entre *appelons* et *appellent*. Entre ces deux ensembles de relations, la régularité formelle n'est que partielle (la régularité quant aux traits sémantiques est, elle, respectée). Nous reviendrons aux chapitres 2 et plus encore 4 sur les façons dont on peut distinguer les relations « similaires » des relations qui ne le sont pas assez, et les concepts de niveau morphologique que l'on peut en inférer. Nous nous contenterons ainsi à ce stade de définir l'existence d'un type d'unités élémentaires qui sont les unités pertinentes pour définir de telles relations. Nous les qualifierons de *formes fléchies* ou simplement *formes*. Nous dirons qu'une forme *réalise* ou *exprime* les traits sémantiques correspondant, tels qu'identifiés au travers des relations décrites ci-dessus.

Le cas le plus simple est celui où, dans un énoncé donné, une forme est également un mot morphosyntaxique unique et un mot sémantique unique [25]. Une telle forme est qualifié de *forme simple* (exemple : *clémentines*). Lorsqu'une forme est également un mot sémantique unique mais correspond à plusieurs mots morphosyntaxiques, elle est qualifiée de *forme périphrastique* (exemple : *ai mangé*) [26]. À l'inverse, lorsqu'un mot sémantique est composé de plusieurs mots morphosyntaxiques qui, chacun, sont des formes, nous parlerons de *composé* (exemple : *pommes de terre*). Dans un tel cas, on notera qu'il y a non-correspondance entre le récipiendaire légitime des traits sémantiques (le composé) et le niveau où ces traits sémantiques sont exprimés (les formes qui constituent le composé). Cela conduit à réviser la définition ci-dessus en introduisant la notion de *trait morphologique* : prototypiquement, traits morphologiques et traits sémantiques coïncident quant à leur récipiendaire et quant à leurs valeurs. Mais deux cas de non-correspondance peuvent se produire. Le premier cas est celui des cas de non-correspondance quant aux valeurs. C'est le cas des verbes dits « déponents » du latin, qui font usage de formes généralement passives (trait morphologique) pour exprimer l'actif (trait sémantique). Le second cas est justement celui des composés tels que *pommes de terre*, pour lesquelles les valeurs correspondent, mais les unités qui expriment les deux types de traits diffèrent (une forme, ici *pommes*, pour les traits morphologiques ; un mot sémantique, ici *pommes de terre*, pour les traits sémantiques).

Nous qualifions de *mot-forme* toute forme qui n'est pas une forme périphrastique. Par abus de langage, dans le reste de ce document, nous utiliserons le terme de *forme* (et même de *forme fléchie*) comme synonyme de *mot-forme*. Autrement dit, dans la suite de ce document, les formes périphrastiques ne sont pas des formes.

Un ensemble de formes (non périphrastiques, donc) ne se distinguant que par les traits morphologiques exprimés est qualifié de *lemme*. Par abus de langage, on pourra également appeler lemme l'une des formes composant ce lemme, dite *forme de citation* choisie selon

25. Naturellement, il arrive qu'une même forme soit susceptible de correspondre à différents mots sémantiques. Généralement, le contexte permet de désambiguïser cette correspondance multiple.

26. On notera qu'une forme périphrastique n'est pas nécessairement contiguë (cf. *j'ai tout mangé*).

un critère généralement arbitraire mais consensuel. Par exemple, il est d'usage de dénoter un lemme verbal en français par son infinitif. On pourra noter que cela n'est pas sans créer des ambiguïtés : ainsi, la forme infinitive *ressortir* fait partie de deux lemmes bien distincts, qui diffèrent par exemple par leur forme de première personne du pluriel du présent de l'indicatif (*nous ressortons* vs. *nous ressortissons*).

En première approximation — mais nous y reviendrons longuement dans les chapitres suivants — on peut définir la notion de *classe flexionnelle* comme suit : une classe flexionnelle est un ensemble de lemmes exhibant un nombre important de régularités dont les relations entre formes. Le sens à donner à l'adjectif « important » est ici laissé volontairement flou.

Un lexique morphologique est alors une collection de lemmes. Chaque lemme peut être exprimé par la liste de ses formes et des traits morphologiques et/ou sémantiques correspondant (nous parlerons de *lexique extensionnel*) ou par une forme de citation et les informations nécessaires pour reconstruire ses formes (notamment mais pas uniquement un identifiant de classe flexionnelle, pour peu que cet identifiant soit défini par ailleurs).

1.3.6 Propriétés phonologiques

Dans ce document, nous ne traiterons pas des aspects phonétiques, phonologiques et prosodiques de la langue. Nous nous contenterons donc de renvoyer ici par exemple à Hall et Kleinhenz (1999), où la notion de *mot phonologique*, terme proposé initialement par Dixon (1977) est étudiée en détails, ou à Dixon et Aikhenvald (2003, p. 13ff.) pour une discussion plus générique. Elle y est définie, pour simplifier, comme le récipiendaire légitime d'un certain nombre de propriétés qui s'appliquent à des unités de taille intermédiaire entre la syllabe et le constituant phonologique ou prosodique : « il constitue le domaine pour de multiples généralisations phonologiques » [27] (Hall et Kleinhenz, 1999, p. 2). Comme l'indique T. A. Hall dans le chapitre introductif de cet ouvrage (p. 2), « tous les auteurs [y] montrent que le mot phonologique n'est pas isomorphe avec le mot grammatical » [28], [29] La propriété phonologique la plus remarquable qu'il est légitime d'assigner à l'unité qualifiée de mot phonologique est l'accent principal. Le cas de décalage le plus fréquent est celui formé par deux mots syntaxiques (correspondant typiquement à deux mots-formes) qui ne sont porteur que d'un seul accent principal. Dans de tels cas, l'un des deux mots syntaxiques est souvent dénué d'autonomie phonologique, au sens où il n'existe pas de contextes où il peut porter un accent. On le qualifie alors de *clitique*.

27. *[…] it forms the domain for various phonological generalizations*
28. *[…] all of the authors show that the pword is non-isomorphic with the grammatical word.*
29. Le terme de « mot grammatical » est ici à interpréter, en première approximation, comme une approximation jointe de la notion de mot-forme et de celle de mot morphosyntaxique telles que définies plus haut.

Nous en resterons là pour cet aperçu des questions liées aux propriétés phonologiques, au mot phonologique, et aux autres domaines phonologiques et phonétiques constituant l'ensemble de la hiérarchie phonologique. Ces sujets, complexes et que nous n'avons pour ainsi dire même pas évoqués ici, ne sont pas abordés dans notre travail.

1.3.7 Propriétés typographiques

Traiter de données textuelles consiste à traiter, avant tout, des flux de caractères encodant des énoncés selon l'un des *systèmes d'écriture* utilisés à travers le monde. Un système d'écriture encode les énoncés au moyen de séquences de signes choisis parmi un inventaire de signes, les *graphèmes*, qui dépend du système utilisé.

La majorité des systèmes d'écriture actuellement utilisés reposent sur l'un des descendants directs ou indirects de l'alphabet phénicien (les alphabets latin, grec et cyrillique et les alphasyllabaires arabe et hébreu, notamment) et font usage de l'espace comme séparateur typographique [30]. C'est également le cas d'autres systèmes d'écriture, mais d'autres caractères que l'espace sont parfois utilisés, et notamment le double point [31]. On peut alors définir une unité typographique, *mot typographique* ou *token*, de la façon suivante : un token est une séquence contiguë de caractères délimités de part et d'autre par un séparateur typographique ou par un signe de ponctuation ; par ailleurs, un signe de ponctuation (ou, dans certaines définitions, une séquence contiguë de signes de ponctuation) est token en soi. Cette définition classique du token, pour être applicable, nécessite de se doter d'un inventaire de symboles de ponctuation, éventuellement complété par des conventions contextuelles permettant de distinguer les cas où un même caractère est une ponctuation et ceux où ce n'est pas le cas. Ainsi, en français, il est souvent fait recours à de telles conventions contextuelles pour le trait-d'union et l'apostrophe, qui sont parfois des symboles de ponctuation mais gagnent parfois à être considérés comme faisant partie d'un token plus large, qu'ils séparent parfois du token précédent ou suivant (cf. *aujourd'hui*, *l'idée* et *'glose'*, ou encore *dors-tu*,

30. L'utilisation généralisée de l'espace comme séparateur typographique dans les textes faisant usage de l'alphabet latin n'est acquise qu'à la fin du premier millénaire de notre ère, après que des moines irlandais aient introduit cette pratique au VII^{ème} siècle. Cette pratique fait suite à plusieurs siècles au cours desquels aucun séparateur n'était généralement utilisé pour transcrire le latin, pratique dite de *scriptio continua*. À la période classique, cependant, les multiples alphabets utilisés autour de la Méditerrannée sont généralement utilisés avec un séparateur typographique, souvent le point-en-haut « · » (grec, latin), parfois un double point « : » (grec), une petite barre verticale (grec mycénien écrit en linéaire B, la plupart des systèmes d'écriture cunéiformes) ou un autre symbole.

31. Symbole « ፡ » de l'alphasyllabaire guèze (sémitique, Éthiopie, éteinte), double-point en carien (anatolien, indo-européen, Anatolie du sud-ouest, éteinte), etc.

scénario-catastrophe et *arc-en-ciel*) [32], [33]. La plupart des systèmes d'écriture faisant usage de séparateurs typographiques ont pour caractéristique que les tokens correspondent, en première approximation et dans la majorité des cas, à des formes. C'est utile dans la mesure où cela permet de procéder de façon approchée au découpage d'un énoncé en formes en appliquant un simple algorithme de tokenisation. Toutefois, comme suggéré plus haut, cette approximation est assez mauvaise pour la plupart des langues concernées, parfois trop. Nous y reviendrons au chapitre 7.

De nombreux systèmes d'écriture ont été ou sont encore utilisés qui, contrairement à ceux que nous venons d'évoquer, ne comportent pas de séparateur délimitant des mots typographiques [34]. C'est aujourd'hui le cas de plusieurs alphasyllabaires (ou *abugidas*) utilisés en Asie du sud-est pour transcrire le thaï, le birman, le khmer, le javanais, le balinais, le soundanais et le laotien, ainsi que des systèmes reposant sur les caractères chinois, ou *sinogrammes*, et utilisés notamment pour transcrire les langues sinitiques et le japonais [35]. D'autres systèmes d'écriture sont à considérer de la même façon, quand bien même ils emploient des séparateurs typographiques, dès lors que ces derniers ne

32. Même dans des langues faisant usage de séparateurs typographiques, on trouve parfois, notamment dans les productions des internautes, des séquences dénuées de séparateur ou utilisant comme séparateur le changement de casse. Le contexte d'utilisation le plus fréquent de cette pratique est celui des langages de programmation, où elle est appelée « *camel case* », mais elle est utilisée depuis longtemps pour former des noms de marques (ainsi *OpenOffice*) ou des quasi-acronymes, et s'emploie de plus en plus généralement, notamment il y a quelques années dans les SMS et désormais sur Twitter, en raison des limitations de longueur. Voici un tweet en français qui illustre cette pratique : *#CamelCase. TiensEncoreUnTermeQueJeNeConnaissaitPas. MêmeSiJePratiqueDepuisLongtemps.*. Un autre exemple, en tchèque, lui aussi issu de Twitter : *milujuTakhleDlouzeNapsanyNazvyPromenychZeVsehoNejvicNaSvete* 'jAimeLesLongsNomsDeVariableÉcritsCommeCeciPlusQueToutAuMonde'. De tels tokens uniques sont naturellement délicats à segmenter en formes, d'autant que les majuscules ne sont pas toujours placées de façon aussi cohérente que dans les exemples donnés ici.

33. Nous ne ferons que mentionner ici un phénomène supplémentaire qui a un impact sur la notion de token, celui des systèmes d'écriture faisant usage de ligatures de façon obligatoire, soit que la forme d'une lettre dépende de ce qui la suit et la précède, soit que plusieurs lettres consécutives ont une forme qui n'est pas identique à la juxtaposition de leurs formes standard (ainsi par exemple l'alphabet arabe et ses dérivés). Tous ces systèmes utilisent l'espace comme séparateur typographique, mais un autre type de séparation est parfois également employé, qui consiste sous certaines conditions à ne pas appliquer une ligature pourtant généralement obligatoire. Les conventions d'usage de cette pratique, conventions naturellement spécifiques à chaque langue, indiquent les cas où elle doit être utilisée pour séparer par exemple un morphe ou un clitique ou pour matérialiser une frontière entres bases formant un composé (cf. persan آن ها (*ân·hâ*) 'ils' et non *آنهـا (*ânhâ*) (cf. Sagot et Walther, 2010b), ou allemand *Auflösung* 'dissolution, dissipation' et non **Auflösung*).

34. C'était déjà le cas en égyptien hiéroglyphique, à ceci près que des déterminatifs sont placés à la fin du mot dont ils précisent le champ sémantique, fournissant ainsi explicitement la position de certaines frontières de mots. Comme mentionné plus haut, la pratique dite de *scriptio continua* consistant à ne pas utiliser de séparateur entre mots apparaît en latin à l'époque post-classique, suivant ainsi une pratique alors en vigueur pour écrire le grec. Elle n'a disparu que progressivement, à la fin du premier millénaire. Elle est traditionnellement la norme pour les alphasyllabaires brahmiques de la sphère indienne et le système d'écriture hangeul utilisé pour transcrire le coréen, mais l'espace est désormais de plus en plus utilisé dans la plupart des premiers et presque systématiquement avec le hangeul.

35. Historiquement, les sinogrammes ont servi à transcrire de nombreuses autres langues dont beaucoup ont changé de système d'écriture depuis, soit en remplaçant totalement l'écriture sinitique par un autre système (vietnamien), soit en conservant une proportion variable de sinogrammes au sein d'un système différent (kanjis en japonais, hanji en coréen, caractères chinois dans le système sawndip utilisé pour

séparent pas des unités se rapprochant même de loin de la notion intuitive de mot, mais séparent d'autres types d'unités, et notamment des syllabes. C'est le cas dans le système d'écriture utilisé actuellement pour transcrire le vietnamien [36] et dans le système d'écriture tibétain [37]. Dans de tels systèmes, seules les marques de ponctuation indiquent des frontières entre formes, l'espace pouvant alors jouer le rôle de ponctuation faible [38].

En l'absence de séparateur typographique permettant un découpage en unité approchant la notion de forme, on peut définir la notion de token de façon simple en lui conservant son double caractère formel et déterministe. Pour les systèmes alphabétiques ou alphasyllabiques sans séparateur typographique pertinent, le plus économique est souvent de conserver la définition classique du token, malgré l'absence de séparateur typographique : un token est alors soit un symbole de ponctuation soit une séquence de symboles autres que des ponctuations (ou des espaces, par exemple en vietnamien) et délimités par des ponctuations (ou des espaces). Chaque token recouvre alors le plus souvent plusieurs formes. Pour les systèmes d'écriture comparables au système chinois, il est plus adapté de considérer chaque caractère (sinogramme ou autre) comme un token, de même que l'on assimilera la transcription de chacune des syllabes en vietnamien ou en tibétain comme autant de tokens, puisqu'ils sont typographiquement isolés. L'identification des formes pourra alors être assimilée à une tâche de regroupement des tokens en formes, c'est-à-dire à une tâche s'apparentant à la reconnaissance des mots composés dans les langues occidentales (Magistry, 2013, ch. 4) [39]. Nous y reviendrons au chapitre 7 (section 7.4).

transcrire les langues zhuang, langues de la plus importante minorité non-han de Chine et apparentées au thaï et au lao).

36. Système alphabétique reposant sur l'alphabet latin étendu par de nombreux diactitiques et où le séparateur entre syllabes est l'espace.

37. Le séparateur y est le *tsheg*, un signe spécial : « ་ ».

38. Sauf bien sûr dans les systèmes, comme le système vietnamien, où il est un séparateur de syllabes.

39. Des parallèles peut-être plus pertinents entre regroupement des sinogrammes en formes et interface tokens-formes en français pourraient être cités, tels que le contraste entre *betterave* (PL *betteraves*) et *chou-rave* (PL *choux-raves*), tous deux formés à partir d'un premier nom (BETTE, CHOU) et d'un second nom RAVE (< Anc. Fr. RABE < Fr-Prov. RABA < Lat. *rapa*, PL de RAPŪM 'navet'). On y voit que le degré d'intégration des deux composants a des conséquences opposées sur la typographie et la morphologie, quand bien même le procédé de composition et le deuxième membre du composé sont strictement identiques. Un autre exemple est le verbe récent COPIER-COLLER, qui est l'un des rares cas en français contemporain de composé verbe-verbe (avec notamment le très similaire COUPER-COLLER). On notera que sa conjugaison reflète ici encore une incertitude quant au degré d'intégration : on semble trouver principalement *copier-coller* (INF), *copie-colle* (IND.PRS.1S), *copié-collé* (PST.PTCP), *copiais-collais* (IND.IPFV.PST.1S) mais *copie-collerai* (IND.FUT.1S) ou *copie-collerais* (COND.PRS.1S), comme si le nombre de syllabes du matériau flexionnel suffixal déterminait la flexibilité du premier terme du composé. On pourrait réfléchir à mettre un tel exemple avec les composés verbe-verbe en mandarin, qui sont fréquents (exemple : 看見 'voir, apercevoir', littéralement approximativement 'voir-(a)percevoir').

1.4 Bilan

Nous avons passé en revue, de façon assez superficielle, un certain nombre de propriétés linguistiques que l'on peut associer à des unités élementaires résultat d'une segmentation d'un énoncé. Nous en avons tiré des ébauches de définition de plusieurs types d'unités élémentaires. Le cas le plus simple est naturellement celui où une même unité élémentaire est la récipiendaire légitime des différents types de propriétés. C'est par exemple le cas de *clémentines*, qui est à la fois, entre autres, un terme, un mot sémantique, un mot syntaxique, un mot morphosyntaxique, une forme fléchie et un token. Mais les cas de non-correspondance, dont nous avons cité quelques uns, sont nombreux. Ils sont associés à plusieurs des problématiques classiques en traitement automatique des langues, et peut-être plus généralement en linguistique : mots composés, amalgames, locutions et expressions idiomatiques, clitiques, formes périphrastiques, etc.

Comme indiqué au début de ce chapitre, ce rapide tour d'horizon n'avait pas pour ambition de proposer une étude linguistique approfondie de la notion d'unité linguistique ou lexicale élementaire. Du reste, la pertinence même d'une telle notion est ici avant tout ancrée par son utilisation généralisée en traitement automatique des langues. Toutefois, différents types d'unités seront manipulés tout au long de ce document, et nous tenions à décrire voire définir, ne serait-ce que de façon supercifielle et approchée, ces différents types d'unités et leurs possibles non-correspondances.

Chapitre 2

Le lexique morphologique : modélisation et implémentation

Sommaire

L'analyse morphologique flexionnelle est l'une des tâches les plus élémentaires de tout système de traitement automatique des langues, dans la mesure où elle constitue souvent l'interface indispensable entre les mots et les représentations abstraites que peuvent constituer des structures ou des propriétés syntaxiques, sémantiques ou discursives, tant en analyse qu'en génération. Computationnellement, la façon la plus simple de gérer la morphologie dans un lexique consiste à disposer d'un inventaire de formes fléchies associées à leur lemme, à leur catégorie et aux traits morphosyntaxiques qu'elles expriment. Un tel lexique, que nous qualifierons de *lexique morphologique extensionnel*, peut en effet être encodé de façon efficace, notamment sous forme d'automates finis ou, plus simplement, de tables de hachage.

Une telle approche est néanmoins réductrice et limitée. Tout d'abord, faire l'inventaire de toutes les formes, c'est-à-dire construire un *lexique morphologique extensionnel*, n'a de sens que pour des langues dont la morphologie est d'une richesse raisonnable, telles que les principales langues ouest-européennes et plus généralement la majorité des langues classées traditionnellement comme fusionnelles. Dès lors que le système morphologique d'une langue est très riche, comme c'est le cas par exemple, mais pas seulement, pour de nombreuses langues agglutinantes [1] ou, *a fortiori*, polysynthétiques [2], dresser l'inventaire de toutes les formes fléchies devient inutilement complexe, sinon impossible en pratique : un même lemme peut avoir des milliers de formes fléchies, et les mécanismes dérivationnels productifs peuvent induire des millions de lemmes distincts [3, 4]. Il est alors nécessaire de développer des modèles de la façon dont les formes sont construites que l'on peut exploiter à la volée.

Ensuite, le caractère énumératif d'un lexique morphologique extensionnel en fait un inventaire très redondant qui masque le haut degré de régularité des systèmes morphologiques. D'un point de vue linguistique, il est légitime de chercher à extraire de ces régularités des généralisations pertinentes et de les représenter au sein d'un modèle formel de la morphologie, qui intègre des informations lexicales et une *grammaire morphologique* : on retrouve la même problématique qu'au paragraphe précédent, celle du dépassement du simple inventaire de formes fléchies, mais ici sous un angle différent. Naturellement, de nombreux modèles formels de la morphologie ont été proposés, qui reposent sur des principes différents et peuvent avoir des motivations distinctes (linguistiques et/ou TAL, notamment). Nous en proposons un aperçu à la section A.1.

1. Rappelons qu'une langue agglutinante « prototypique » est caractérisée par le fait que les traits morphosyntaxiques exprimés par la morphologie le sont au moyen de morphes qui, chacun, correspondent à l'un de ces traits de façon biunivoque ou quasiment biunivoque (des variations peuvent être causées notamment par des phénomènes telles que l'harmonie vocalique). Quelques exemples : le turc, le japonais, le finnois, le hongrois, le tamoul ou le basque.

2. Une langue polysynthétique est caractérisée par l'utilisation de séquences de morphes dont aucun ou presque ne peut exister par lui-même, mais dont le sens est complexe (impliquant typiquement plusieurs morphes lexicaux, c'est-à-dire qui expriment un prédicat ou une entité). Un exemple est fourni à la note 4 ci-dessous. Exemples : les variétés de groenlandais, le mohawk (et un certain nombre d'autres langues des Amériques), le tchouktche, et plusieurs langues d'Australie et de Papouasie.

3. L'un n'étant pas incompatible avec l'autre, d'autant que la frontière entre flexion et dérivation, délicate à caractériser empiriquement, peut être encore moins nette dans ce type de systèmes morphologiques. Ceci dit, une étude approfondie de la distinction entre flexion et dérivation dépasse largement le cadre de ce document. Nous nous contenterons de faire l'hypothèse que l'on peut poser l'existence d'une distinction entre ces deux phénomènes.

4. Pour le cas du turc, voir par exemple Hankamer (1989) cité par Hakkani-Tür *et al.* (2002). On peut également citer l'exemple de Fortescue (1999) cité par Ackerman et Malouf (2013) : en kalaallisut (ou groenlandais occidental, inuit, eskimo-aléoute), la forme unique *aju-nngit-su-liur-vigi-nnit-tuar-tu-u-nngil-aq*, qui se glose *BE.GOOD*-NEG-PART-*MAKE-HAVE.AS.PLACE.OF*-ANTIP-*ALL.THE.TIME*-PART-*BE*-NEG-3SG.IND, se traduit par 'ce n'est pas (vraiment) un bienfaiteur'. Un autre exemple classique est fourni par Kibrik (2001), qui indique qu'en artchi (lezghique, nakho-daghestanien), le nombre total de formes pour un paradigme verbal est de 1 502 839.

Mais représenter de façon linguistiquement pertinente le système morphologique d'une langue ne saurait se contenter d'une approche purement extensionnelle [5].

Enfin, une approche purement extensionnelle ne couvre pas à elle seule les innovations lexicales. S'il est possible de développer des systèmes qui s'appuieraient sur un lexique extensionnel pour analyser une forme inconnue, par exemple en s'appuyant sur la notion d'analogie, de tels systèmes ne peuvent que s'appuyer sur un certain nombre d'hypothèses sur la structure même des formes fléchies, complétant ainsi le lexique extensionnel d'informations supplémentaires, souvent implicites, sur le système morphologique considéré. Nous renvoyons à la section A.1.3 pour une discussion plus approfondie sur ce sujet.

Le développement de notre architecture lexicale morphologique et syntaxique Alexina a ainsi naturellement impliqué celui d'un formalisme morphologique permettant de représenter à la fois les propriétés lexico-morphologiques des unités lexicales composant un lexique et une grammaire morphologique associée. Ce travail a été réalisé en deux étapes. Dans un premier temps, nous avons développé ce que nous appellerons le formalisme morphologique Alexina d'origine et noterons $\text{Alexina}_{\text{morph}}$, initialement dans le cadre du lexique Le*fff* pour le français (Clément *et al.*, 2004 ; Sagot *et al.*, 2006) puis en l'enrichissant dans le cadre de travaux sur l'espagnol, le slovaque le polonais et le persan (Sagot, 2005a ; Sagot *et al.*, 2006 ; Sagot, 2006, 2007 ; Molinero *et al.*, 2009b ; Sagot, 2010 ; Sagot et Walther, 2010b). Mais ces travaux ont également montré certaines limites d'$\text{Alexina}_{\text{morph}}$. En conséquence, et dans un deuxième temps, nous avons travaillé à la refonte de ce formalisme, en interaction avec les travaux en morphologie formelle de Walther (2011b, 2013b, 2016). Ces travaux, qui ont conduit au modèle formel PARSLI de la morphologie flexionnelle (PARadigm Shape and Lexicon Interface), ont permis le développement d'un nouveau formalisme morphologique plus complet et plus pertinent sur les plans théoriques et morphologiques, appelé $\text{Alexina}_{\text{PARSLI}}$. Ce dernier met en œuvre des concepts issus de PARSLI, hérite de propriétés d'$\text{Alexina}_{\text{morph}}$, et intègre des améliorations propres.

2.1 $\text{Alexina}_{\text{morph}}$ [6]

$\text{Alexina}_{\text{morph}}$, le premier formalisme morphologique d'Alexina (cf. notamment Sagot, 2005a), avait déjà fait ses preuves pour l'implémentation de grammaires flexionnelles

5. On peut naturellement aussi extraire les mots les plus fréquents d'un corpus donné, le considérer comme un lexique extensionnel (incomplet), et disposer de mécanismes de gestion des mots inconnus. Mais de tels mécanismes ne sont alors pas très éloignés de systèmes de flexion automatique.

6. Le travail décrit dans cette section a été initié dans le cadre du développement du Le*fff* (Clément *et al.*, 2004 ; Sagot *et al.*, 2006 ; Sagot, 2010), puis a évolué fortement à l'occasion de travaux sur le slovaque (Sagot, 2005a) puis le polonais (Sagot, 2007). Depuis lors, il n'a cessé d'être utilisé dans plusieurs de nos lexiques, malgré le développement, en parallèle du formalisme plus avancé, $\text{Alexina}_{\text{PARSLI}}$, évoqué à la section suivante.

de différentes langues associées à des lexiques à large ou moyenne couverture (cf. tableau 2.1). Le premier de ces lexiques a été le Le*fff*, lexique morphologique et syntaxique à large couverture du français, librement disponible et largement utilisé dans la communauté du traitement automatique des langues francophone (Sagot *et al.*, 2006 ; Sagot, 2010). Nous reviendrons brièvement à la section A.6 sur l'historique du développement du Le*fff* et sur les utilisations qui en sont faites, lorsque nous décrirons le développement de son volet syntaxique. Les autres lexiques morphologiques Alexina dont le développement a fortement contribué à l'amélioration d'Alexina$_{\text{morph}}$ sont notamment le lexique morphologique du slovaque SkLex (Sagot, 2005a) et le lexique morphologique du polonais PolLex (Sagot, 2007).

Le développement de ces lexiques a été grandement facilité par différentes techniques de maintenance de lexiques morphologiques, d'extraction automatique d'informations lexico-morphologiques, dont certaines font l'objet du chapitre suivant, ainsi que d'interfaces de validation associées. De plus, tous les lexiques Alexina, à l'image du Le*fff*, sont librement disponibles, ce qui garantit que les analyses morphologiques et les données lexicales peuvent être vérifiées et utilisées par tous, que ce soit à des fins typologiques, morphologiques ou de traitement automatique des langues.

Lexique	Langue	#lemmes	#lexèmes	#formes	#formes dist.	Références
Le*fff* [7]	français	120,000	125,000	550,000	460,000	(Sagot, 2010 ; Walther et Sagot, 2011a)
Le*ff*e	espagnol	180,000	180,000	1,500,000	700,000	(Molinero *et al.*, 2009b)
Le*ff*ga	galicien	70,000	70,000	750,000	500,000	(Molinero *et al.*, 2009b)
Le*ff*la	latin	2,200	2,200	115,000	96,000	(Walther, 2013b)
EnLex	anglais	350,000	350,000	580,000	510,000	(Sagot, 2010)
DeLex	allemand	63,000	63,000	2,100,000	405,000	(Sagot, 2014)
PolLex	polonais	240,000	240,000	1,400,000	360,000	(Sagot, 2007)
SkLex	slovaque	50,000	50,000	470,000	250,000	(Sagot, 2005a)
PerLex	persan	30,000	30,000	550,000	460,000	(Sagot et Walther, 2010b ; Sagot *et al.*, 2011c)
KurLex	kurmanji	22,000	22,000	410,000	240,000	(Walther *et al.*, 2010)
SoraLex	sorani	520	520	30,000	25,000	(Walther et Sagot, 2010)
MaltLex	maltais	560	560	9,000	7,200	(Camilleri et Walther, 2012)
KhaLex	khaling	591	744	170,000	54,000	(Walther *et al.*, 2013)

Tableau 2.1 – Lexiques Alexina. Les lignes grisées correspondent à des lexiques Alexina$_{\text{PARSLI}}$.

La façon dont Alexina encode la morphologie repose explicitement sur une approche paradigmatique. Chaque entrée lexicale, qui correspond en principe à un lexème (mais parfois à un lemme), est associée à une classe flexionnelle, comme illustré dans la partie

7. La version Alexina$_{\text{PARSLI}}$ de la grammaire morphologique du Le*fff* est celle appelée new dans (Sagot et Walther, 2011). La version antérieure, qui est encore la version « officielle », repose quant à elle sur le formalisme Alexina d'origine (Sagot, 2010). Par ailleurs, les données quantiatives fournies ici correspondent à la version 3.3 du Le*fff*.

supérieure de la figure 2.1 par cinq entrées lexicales du Le*fff* [8]. Chaque *entrée intensionnelle* est composée d'une forme de citation et d'une *classe flexionnelle*. Dans la grammaire morphologique, chaque classe flexionnelle est définie explicitement à l'aide de règles de réalisation qui décrivent comment elle construit les paradigmes. Pour certaines entrées lexicales, la classe flexionnelle est associée avec une *variante* de classe flexionnelle (il pourrait y en avoir plusieurs) qui permettent de choisir des règles spécifiques pour produire certaines des formes du paradigmes (on pourrait qualifier ces variantes de *diacritiques de classes flexionnelles*). Par exemple, dans la figure 2.1, les variantes *dbl* et *std* concernent respectivement les verbes du premier groupe (classe flexionnelle *v-er*) qui doublent leur consonne de fin de radical dans certaines cases (cf. *appeler* / *appelle*, *jeter* / *jette*) et les verbes du premier groupe qui n'ont pas ce comportement (cf. *peler* / *pèle*, *acheter* / *achète*). La partie inférieure de la figure 2.1 montre quelques entrées fléchies, ou *entrées extensionnelles*, produites par les entrées lexicales de la partie supérieure de la figure au moyen de la grammaire morphologique. Pour chaque entrée extensionnelle, la catégorie morphologique du lexème concerné est indiquée à côté de la forme fléchie, ainsi que la forme de citation et l'étiquette morphologique qui encode les structures de traits morphosyntaxiques exprimés [9].

Une grammaire morphologique Alexina comporte deux sections principales :

— un ensemble de règles morphophonologiques, ou, plus précisément, de règles morphographémiques simulant des règles morphophonologiques, tous les lexiques Alexina existant à ce jour étant des lexiques orthographiques ; [10] dans la suite, nous utiliserons improprement le terme de « règle morphophonologique » pour dénoter ces règles, quand bien même elles sont morphographémiques [11] ;

— la section morphologique proprement dite, c'est-à-dire la définition de chacune des classes flexionnelles.

La section morphologique proprement dite, dans une grammaire morphologique écrite dans le formalisme originel d'Alexina, définit les classes flexionnelles au moyen de règles de réalisation des formes, chacune étant associé avec l'étiquette

8. Pour simplifier, les informations syntaxiques ne sont pas indiquées ici. Nous y reviendrons notamment au chapitre 5.

9. Les classes flexionnelles mentionnées dans la figure 2.1 sont *v-er* pour la classe régulière et productive des verbes dits du premier groupe, *v-ir2* pour la classe régulière mais à peu près non productive des verbes dits du deuxième groupe, *v-ir3* pour les verbes du troisième groupe en *-ir* et *v55* pour l'une des classes irrégulières de verbes du troisième groupe. Les étiquettes associées aux formes fléchies en quatrième colonne utilisent un format directement inspiré du projet MULTEXT. Par exemple, « P3s » s'interprète comme indicatif présent (P), 3ème personne (3) du singulier (s). « J » désigne l'indicatif passé simple, et « G » le participe présent.

10. La section morphophonologique commence généralement par la définition de graphèmes (y compris des digraphes voire des trigraphes) et de classes de graphèmes, qui correspondent souvent à des classes de phonèmes (par exemple, les voyelles avant). Ces classes peuvent alors être utilisées dans la définition de règles morphographémiques, mais également pour énoncer des conditions sur la compatibilité des règles réalisationnelles voire des classes flexionnelles elles-mêmes avec un radical donné.

11. La différence n'est toutefois pas toujours si nette, notamment dans des langues comme le slovaque où la correspondance entre graphèmes et phonèmes est presque parfaite.

```
accoutumer  v-er:std
appeler     v-er:dbl
enrichir    v-ir2
dormir      v-ir3
admettre    v55
```

```
accoutuma    v    accoutumer   J3s
accoutume    v    accoutumer   PS13s
accoutumant  v    accoutumer    G
appela       v    appeler      J3s
appelle      v    appeler      PS13s
appelant     v    appeler      G
enrichit     v    enrichir     J3s
enrichit     v    enrichir     P3s
enrichissant v    enrichir     G
dormit       v    dormir       J3s
dort         v    dormir       P3s
dormant      v    dormir       G
admit        v    admettre     J3
admet        v    admettre     P3
admettant    v    admettre     G
```

FIGURE 2.1 – Entrées lexicales du Le*fff* (niveau morphologique uniquement) : au dessus, quelques entrées intensionnelles ; en dessous, quelques unes des entrées extensionnelles correspondantes.

morphophonologique qu'elle permet de réaliser. Ces règles peuvent faire usage d'opérations de suffixation et/ou de préfixation. Tout autre opération réalisationnelle (par exemple, une alternance vocalique) doit être simulée en deux étapes, via une règle d'affixation qui insère l'information nécessaire pour que, dans un deuxième temps, une règle morphophonologique artificielle s'applique et réalise effectivement l'opération nécessaire [12]. Une règle de réalisation peut également ne pas être explicite mais prendre la forme d'une règle d'héritage, c'est-à-dire indiquer qu'une étiquette morphologique sera réalisée par la même forme qu'une autre : il s'agit de règles de renvoi au sens de Zwicky (1985). Ceci constitue une modélisation simple (et directionnelle) de la notion de syncrétisme [13]. Une classe flexionnelle peut également hériter d'une autre classe flexionnelle, soit en totalité soit en partie, puis spécifier des règles de réalisation pour certaines étiquettes morphologiques qui remplaceront celles de la classe dont elle a hérité. Ainsi, dans le Le*fff*, la classe flexionnelle *adj-4* des adjectifs qui se fléchissent en

12. Par exemple, en français, produire *appelle* à partir d'*appeler* et *jette* à partir de *jeter* de façon unifiée nécessite la mise en place d'une opération non concaténative, à savoir la duplication de la consonne qui termine le radical. La grammaire standard du Le*fff* utilise ainsi un affixe comme *-2e* suivi d'une règle morphophonologique artificielle qui réécrit *t_2* en *tt_* et *l_2* en *ll_* (« _ » indiquant une frontière entre morphes).

13. Une forme est syncrétique si elle occupe plusieurs cases d'un même paradigme. Par exemple, dans le paradigme verbal du verbe français MANGER, la forme *mange* est un exemple de syncrétisme, puisqu'elle correspond à la fois à la première et à la troisième personne du présent de l'indicatif et du subjonctif, ainsi qu'à la deuxième personne de l'impératif présent.

genre (*-e* au féminin) et en nombre (*-s* au pluriel) hérite de façon globale de la classe flexionnelle *nc-4* des noms communs qui se fléchissent en genre et en nombre de façon identique, comme *doctorant(e)(s)* (cf. figure 2.2) [14]. Par ailleurs, toute règle de réalisation (explicite ou d'héritage) peut contraindre le type d'input sur lequel elle peut s'appliquer, au moyen de *contraintes* positives (`rads=`) ou négatives (`rads_except=`) exprimées sous la forme d'expressions régulières [15]. Enfin, une classe flexionnelle peut également spécifier des règles de construction de lemmes dérivés possibles, là encore par affixation et/ou préfixation, et en spécifiant la classe flexionnelle du lemme dérivé (cf. section 3.3.1) [16].

```
<table name="nc-4" rads="..*[^sxzce]">
  <form suffix="" tag="ms" show="#"/>
  <form suffix="e" tag="fs" show="#"/>
  <form suffix="s" tag="mp" show="#"/>
  <form suffix="es" tag="fp"/>
</table>

<table name="adj-4" rads="..*[^sxzce]">
  <like name="nc-4"/>
</table>
```

FIGURE 2.2 – Tables *nc-4* et *adj-4* dans le Le*fff*. La table *nc-4* est celle des noms se fléchissant en genre (*-e* au féminin) et en nombre (*-s*) au pluriel. On voit que cette table est réservée à des radicaux (et donc des formes de citation, le masculin singulier étant la forme de citation et ayant un suffixe nul) ne se terminant pas en *-s*, *-x*, *-z*, *-c* ou *-e*. La table *adj-4*, table adjectivale assortie de la même contrainte, est définie par héritage complet de la table *nc-4*.

14. Sur le plan technique une grammaire morphologique Alexina est un document XML. Un outil dédié permet de compiler une telle grammaire en différents outils :

- un script de flexion, qui peut fléchir tout lexique intensionnel associé à la grammaire ;
- un outil de « désinflexion », qui permet à partir d'une forme de reconstituer toutes les entrées intensionnelles formellement possibles dont, selon la grammaire, la forme serait une forme fléchie,
- un outil de dérivation, qui produit tous les lexèmes dérivés possibles selon les règles dérivationnelles incluses dans la grammaire.

Nous avons conservé cette architecture technique avec Alexina PNRSLI.

15. Il est par exemple possible d'avoir dans une même classe deux règles pour la même étiquette morphosyntaxique, et de spécifier que la première ne s'applique que sur un radical qui se termine en consonne ou en semi-voyelle et que la seconde ne s'applique que sur un radical qui se termine en voyelle ou en semi-voyelle. Dans ce cas, la case correspondante sera surabondante dans le paradigme de tout lemme dont le radical se termine par une semi-voyelle, mais pour ces lemmes seulement.

16. Naturellement, ces règles dérivationnelles peuvent aussi être assorties de contraintes sur le radical.

2.2 Alexina$_{\text{ParSlI}}$ [17]

Comme indiqué plus haut, Alexina$_{\text{morph}}$ n'était pas pleinement satisfaisant. Tout d'abord, la structure macroscopique du système morphologique n'y était pas représentée de façon satisfaisante. Ensuite, l'utilisation de règles morphographémiques pour encoder les règles de réalisation non-concaténatives a montré ses limites, notamment au cours du développement du lexique PerLex du persan. Nous avons donc cherché à rendre possible l'implémentation d'analyses morphologiques formalisées au sein d'un modèle typologiquement plus pertinent. Une telle ambition a naturellement des répercussions sur le développement de ressources lexicales, y compris pour le traitement automatique des langues en général et les outils tirant parti de telles ressources en particulier, en permettant un développement accéléré de lexiques morphologiques linguistiquement motivés.

L'analyse des différentes approches formelles de la morphologie flexionnelle, dont nous donnons un aperçu à la section A.1, nous a conduit nous tourner vers une approche inférentielle réalisationnelle au sens de Stump (2001), et plus précisément à travailler en lien étroit avec le développement par Walther du modèle formel ParSlI (PARadigm Shape and Lexicon Interface ; Walther, 2011b, 2013b, 2016). Le résultat en est le formalisme Alexina$_{\text{ParSlI}}$, qui est aujourd'hui intégré à Alexina, aux côtés d'Alexina$_{\text{morph}}$, de sorte que l'un ou l'autre de ces deux formalismes peuvent être utilisés pour le développement d'un nouveau lexique Alexina. Alexina$_{\text{ParSlI}}$ a été développé à la fois comme une extension d'Alexina$_{\text{morph}}$ et comme un formalisme d'implémentation [18] pour ParSlI [19].Alexina$_{\text{ParSlI}}$ permet ainsi d'encoder d'une façon computationnellement

17. Le travail décrit dans cette section a été réalisé en grande partie en collaboration avec Géraldine Walther, alors doctorante au Laboratoire de Linguistique Formelle, Université Paris Diderot, sous la direction d'Anne Abeillé (LLF, Université Paris-Diderot) et d'Olivier Bonami (LLF, Université Paris-Sorbonne). Il a fait l'objet de plusieurs publications communes (Sagot et Walther, 2011 ; Walther et Sagot, 2011a ; Sagot et Walther, 2013). Ces travaux reposent sur des travaux théoriques menés par Géraldine Walther (Walther, 2011b, 2013b, 2016) qui sont évoqués brièvement à la section 2.2.1, mais également sur des expériences de développement de lexiques morphologiques à des fins de TAL ou de linguistique quantitative, ces dernières ayant été menées pour certaines en collaboration avec Géraldine Walther et, pour l'une d'entre elles, de Guillaume Jacques (CRLAO, CNRS). Certaines d'entre elles seront évoquées respectivement aux section 3.1.3, 4.1.1 et 4.1.2.

18. Nous utilisons ici le terme d'implémentation pour dénoter le fait qu'Alexina$_{\text{ParSlI}}$ fournit un moyen de créer et de manipuler des ressources électroniques (lexiques, grammaires) qui correspondent à des analyses morphologiques développées au sein du modèle ParSlI de la morphologie flexionnelle. Alexina$_{\text{ParSlI}}$ est à la fois un langage et un ensemble d'outils qui peuvent traiter une description morphologique écrite dans ce langage, par exemple pour en induire un outil de flexion automatique.

19. On peut oser une comparaison avec la formalisation et l'implémentation de la syntaxe : on peut comparer le formalisme morphologique Alexina$_{\text{ParSlI}}$ dont il est question dans ce chapitre avec la plateforme LKB (Copestake, 2002), là où c'est le modèle ParSlI, sur lequel repose Alexina$_{\text{ParSlI}}$, qui correspond au modèle théorique HPSG dont LKB est une implémentation. Pour poursuivre le parallèle, des outils reposant sur les automates finis comme XFST (Beesley et Karttunen, 2003) ou FOMA (Hulden, 2009) correspondent à des générateurs d'analyseurs syntaxiques pour les grammaires non-contextuelles telles que la paire Lex/Yacc (http ://dinosaur.compilertools.net) ou le système Syntax (Boullier et Deschamp, 1988–2007), formalismes syntaxiques purement algébriques qui ne reposent pas sur un modèle linguistique complet.

exploitable des descriptions flexionnelles (lexique et grammaire) en élaborant à partir d'une approche morphologiquement motivée.

2.2.1 Le modèle ᚠᚨᚱᛋᛚ de la morphologie flexionnelle

ᚠᚨᚱᛋᛚ (Walther, 2011b, 2013b, 2016) est un modèle formel de la morphologie flexionnelle de type inférentiel-réalisationnel au sens de Stump (2001) repose sur les hypothèses générales de ce qu'il est convenu de nommer l'approche « mot et paradigme » (Word-and-Paradigm, Hockett, 1954). Ainsi, pour un lexème donné, l'ensemble de ses formes fléchies (ou formes) constitue son *paradigme*. Chacune de ces formes remplit une (parfois plusieurs) *cases* du paradigme, c'est-à-dire qu'elle réalise une (parfois plusieurs) des structures de traits morphosyntaxiques pertinentes pour sa catégorie morphosyntaxique. Par exemple, en français, le paradigme d'un adjectif contient quatre cases, un adjectif réalisant en français les traits de genre (M et F) et de nombre (SG et PL).

ᚠᚨᚱᛋᛚ est avant tout un modèle de l'interface entre la structure du paradigme des lexèmes et la structure de leur entrée lexicale. Il été construit dans le but de proposer une formalisation des concepts développés dans le cadre de la *typologie canonique*, approche de la typologie présentée notamment par Corbett (2003, 2007). L'idée fondamentale de la typologie canonique consiste à définir les propriétés d'un système idéal, qui n'a vocation à être attesté dans aucune langue, mais qui constitue un point de repère par rapport auquel on peut étudier la façon dont un système attesté se différencie. Dans le cas de la flexion, un système canonique est caractérisé par une ambiguïté nulle et une régularité maximale. Pour simplifier, on peut caractériser la flexion canonique comme suit (Corbett, 2007 ; Walther et Sagot, 2011a ; Walther, 2013b) :

- quel que soit le lexème, une même case de paradigme sera construite de la même façon à partir du radical (autrement dit, il n'y a qu'une seule classe flexionnelle) ;
- toutes les formes d'un lexème se construisent à partir du même radical ;
- quel que soit le lexème, deux cases différentes d'un paradigme contiennent des formes différentes ;
- chaque lexème dispose d'un radical qui lui est propre ;
- chacune des cases du paradigme d'un même lexème comporte une et une seule forme.

Ainsi, la régularité d'un système flexionnel canonique est totale, et à une forme correspond de façon non ambiguë un lexème et une case.

Naturellement, les systèmes flexionnels attestés ne sont jamais strictement canoniques [20], et les écarts à la canonicité sont au cœur des approches canoniques de la flexion.

20. Plus précisément, aucun système flexionnel canonique n'a été observé, et il n'est pas attendu que cela ait lieu : la flexion canonique est un point extrémal, pas un prototype.

ᴘᴀʀsʟ modélise explicitement les régularités et irrégularités dans les paradigmes individuels et dans le système flexionnel dans son ensemble, c'est-à-dire les *phénomènes non-canoniques* tels que définis en flexion canonique, et ce au moyen d'informations représentées dans l'entrée lexicale, comme nous le verrons plus bas. Nous illustrerons dans la suite de cette section certains de ces phénomènes, qui nous seront utiles par la suite, notamment au chapitre 4. Nous renvoyons à Walther (2013b) pour un inventaire plus détaillé.

2.2.1.1 L'entrée lexicale dans ᴘᴀʀsʟ

ᴘᴀʀsʟ formalise explicitement la notion d'entrée lexicale flexionnelle. Comme illustré à la figure 2.3, chaque entrée, correspondant à un lemme, est définie au moyen des éléments suivants sur lesquels nous allons revenir ci-dessous : (i) une base phonologique I-PHON qui est le point de départ sur lequel les règles flexionnelles s'appliquent en séquence, (ii) une catégorie flexionnelle I-CAT, (iii) un ensemble MSF de structures de traits morphosyntaxiques exprimables, (iii) un ensemble S-STEM de radicaux supplétifs et un ensemble S-FORM de formes supplétives, (iv) un schème flexionnel I-PAT composé d'un ensemble de sous-schèmes qui spécifient les combinaisons de règles de réalisation à appliquer à parti de la base phonologique pour construire les formes fléchies [21].

BALAYER		
I-PHON	balayer	
I-CAT	*verbe*	
MSF	{ *standard* }	
S-STEM	∅	
S-FORM	∅	
I-PAT	(z^s_{ay}, id),	(z^{exp}_{v1}, id)
	(z^s_{ay}, id),	(z^{exp}_{v1}, id)
	(z^s_{ai}, id),	(z^{exp}_{v1}, id)

FIGURE 2.3 – Entrée lexicale ᴘᴀʀsʟ pour le lemme verbal BALAYER(les z^s et z^{exp} seront explicitées ci-dessous).

Cette représentation des entrées lexicales permet de rendre explicites certains écarts à la canonicité telle qu'esquissée ci-dessus, c'est-à-dire de mettre en avant les différents types

21. Cette manière de présenter la notion de schème comme une indication sur la manière de *produire* ou construire les formes fléchies n'est en rien la manifestation de ce que cette notion ne puisse être utilisée à des fins d'*analyse*. Bien au contraire, l'implémentation en Alexinaᴘᴀʀsʟ d'une description morphologique ᴘᴀʀsʟ peut servir de base au développement d'un analyseur morphologique, par exemple pour proposer tous les lexèmes (inconnus) formellement possibles dont un mot inconnu que l'on cherche à analyser serait une forme fléchie (cf. section 3.1.1).

de phénomènes non-canoniques que l'on peut rencontrer, et dont nous mentionnerons quelques uns dans la suite de cette section. Ainsi, la liste S-STEM permettra d'indiquer les éventuels *radicaux supplétifs* (cf. français *aller* vs. *irai* vs. *vais*) et la liste S-FORM des *formes supplétives* (cf. français *faites* vs. **faisez*). La liste MSF, sauf à utiliser le mot-clé « *standard* » qui indique que toutes les cases attendues au vu de la catégorie morphosyntaxique sont remplies, permet d'indiquer quelles sont les cases du paradigme qui existent effectivement pour l'entrée considérée, ce qui permet le codage explicite des cas de *déficience* (ainsi, pour le dire de façon imagée, le verbe français SEOIR$_{N}$'existe pas à toutes les formes). Le schème flexionnel I-PAT rend également explicite certains phénomènes non-canoniques, comme la surabondance dans le cas de BALAYER. Nous ne rentrons pas ici dans plus de détails, et renvoyons à la suite de cette section et, pour plus d'informations, à (Walther, 2013b) et à (Sagot et Walther, 2013).

2.2.1.2 Structures de traits morphosyntaxiques et catégories flexionnelles

Chaque entrée lexicale est définie par son appartenance à une certaine catégorie flexionnelle. Une catégorie flexionnelle exprime canoniquement un certain ensemble de structures de traits morphosyntaxiques, ensemble défini dans la grammaire [22]. Si un lemme appartient à une catégorie donnée, il exprimera canoniquement l'ensemble de traits morphosyntaxiques correspondant à sa catégorie. Dans ce cas, son ensemble de structures de traits morphosyntaxiques exprimables est noté *standard* dans son entrée lexicale, comme c'est le cas pour BALAYER à la figure 2.3. Certains lemmes, cependant, expriment plus ou moins de structures traits morphosyntaxiques qu'attendu au vu de leur catégorie flexionnelle. Ces écarts seront notés dans le champ MSF de leur entrée lexicale. Le paradigme d'un tel lemme est alors non-canonique : il y a surdifférenciation dès lors qu'il exprime plus de traits qu'attendus [23] et déficience s'il en exprime moins [24].

2.2.1.3 Zones réalisationnelles

L'une des innovations majeures de ᚹᚪᚱᛊᚳᛁ par rapport aux modèles auquel il est comparable (Stump, 2001, 2006 ; Brown et Hippisley, 2012) est la généralisation des

22. En latin, par exemple, un nom peut exprimer deux nombres (SG et PL) et cinq cas (NOM, ACC, GEN, DAT et ABL).

23. Par exemple en exprimant une valeur supplémentaire pour un trait donné, ou en exprimant différemment des valeurs différentes pour un trait généralement non-pertinent pour sa catégorie. C'est par exemple le cas en slovaque où seuls une douzaine de noms disposent d'une forme spécifique pour le vocatif, distincte du nominatif singuler (par exemple *bože !* 'dieu !' vs. *boh* 'dieu').

24. Walther (2013b, p. 222) réserve le terme, parfois utilisé, de *défectivité* pour un phénomène aux conséquences proches mais au mécanisme distinct, causé par la non-spécification de mécanismes de réalisation pour certaines cases du paradigme. C'est le cas du le verbe TRAIRE$_{E}$ n français, qui n'a ni passé simple ni imparfait du subjonctif, faute de façon de construire le radical pour ces cases. C'est très différent du cas de SEOIR, pour lequel de telles règles existent (elles s'appliquent sans problème au verbe ASSEOIR, par exemple), et où le fait que certaines cases ne soient pas remplies doit donc être spécifié au niveau de la liste MSF.

concepts de *partition* de paradigme de Pirelli et Battista (2000) et d'*espace thématique* de Bonami et Boyé (2003) et leur extension à la description de l'exponence (Matthews, 1974), au moyen la notion de *zone réalisationnelle*. Plutôt que d'associer à une entrée lexicale une classe flexionnelle complète, ParSLi l'associe à plusieurs zones réalisationnelles qui, chacune, contiennent un ensemble de règles de réalisation permettant la construction systématique d'un sous-ensemble du paradigme complet pour tout lexème qui en fait usage. Ces ensembles de règles de réalisation peuvent ainsi être combinés de différentes façons pour représenter la manière de réaliser différents types de paradigmes. Ainsi, les paradigmes hétéroclites, dont une illustration en slovaque est donnée à la table 2.2, peuvent être représentés au moyen de la combinaison de zones réalisationnelles généralement utilisées par des lemmes appartenant à différentes classes flexionnelles. Ces données du slovaque montrent comment certains noms d'animaux utilisent ainsi d'une part la zone réalisationnelle couvrant le singulier des noms animés pour construire leurs formes du singulier et d'autre part la zone réalisationnelle couvrant le pluriel des noms inanimés pour construire leurs formes du pluriel. ParSLi définit alors une classe flexionnelle, telle que la classe Z^{exp}_{anim} des noms slovaques animés (cf. table 2.2), comme étant une combinaison significativement fréquente de zones réalisationnelles — cette significativité doit être alors définie, par exemple au moyen de l'application quantitative d'un principe d'économie descriptive à l'échelle de tout le système flexionnel, comme nous le verrons au chapitre 4. La notion de classe flexionnelle est ainsi une notion dérivée de celle de zone réalisationnelle, et n'est pas une primitive du modèle.

L'ensemble des structures de traits morphosyntaxiques que les règles de réalisation d'une zone réalisationnelle expriment définit son ESPACE PARTITIONNANT.

	Z^{exp}_{anim} : MASC. ANIMÉ CHLAP 'garçon, type'		Z^{exp}_{inan} : MASC. INANIMÉ DUB 'chêne'		MASC. HÉTÉROCLITE OROL 'aigle'	
	$z^{exp}_{\text{anim, sg}}$: SG	$z^{exp}_{\text{anim, pl}}$: PL	$z^{exp}_{\text{inan, sg}}$: SG	$z^{exp}_{\text{inan, pl}}$: PL	$z^{exp}_{\text{anim, sg}}$: SG	$z^{exp}_{\text{inan, pl}}$: PL
NOM	*chlap*	*chlap-i*	*dub*	*dub-**y***	*orol*	*orl-**y***
GEN	*chlap-a*	*chlap-ov*	*dub-a*	*dub-ov*	*orl-a*	*orl-ov*
DAT	*chlap-**ovi***	*chlap-om*	*dub-u*	*dub-om*	*orl-**ovi***	*orl-om*
ACC	*chlap-a*	*chlap-ov*	*dub*	*dub-**y***	*orl-a*	*orl-**y***
LOC	*chlap-**ovi***	*chlap-och*	*dub-e*	*dub-och*	*orl-**ovi***	*orl-och*
INS	*chlap-om*	*chlap-mi*	*dub-om*	*dub-mi*	*orl-om*	*orl-ami*

TABLEAU 2.2 – Noms d'animaux hétéroclites en slovaque.

2.2.1.4 Une représentation de la flexion à plusieurs niveaux

ParSLi repose sur une représentation fortement structurée de la réalisation des formes grâce à une représentation faisant usage de plusieurs niveaux réalisationnels. Chaque zone réalisationnelle appartient à un niveau réalisationnel spécifique. Parmi ces

niveaux, le premier est nécessairement un niveau de type radical, qui prend en charge les possibles alternances de radicaux (allomorphie radicale). Mais d'autres niveaux peuvent être définis : des niveaux thématiques et des niveaux d'exponence [25].

La réalisation d'une forme donnée consiste en l'application d'une règle par niveau. Considérons un exemple qui nous sera utile au chapitre 4. En français, la flexion du verbe BALAYER peut être représentée avec un modèle à deux niveaux, un niveau radical et un niveau d'exponence. Une particularité des verbes en *-ayer* tels que BALAYER est qu'ils sont surabondants pour la moitié de leur paradigme : la moitié des cases contient deux formes différentes exprimant les mêmes structures de traits morphosyntaxiques. Ainsi, les formes *balayent* et *balaient* sont toutes deux valides pour exprimer la structure de traits morphosyntaxiques 3PL.PRS.IND, alors que seule la forme *balayons* exprime 1PL.PRS.IND. Plus précisément, chaque case surabondante peut être occupée en partant de deux radicaux distincts, *balay-* et *balai-*, mais avec les mêmes exposants. La façon la plus directe de modéliser cela est de faire usage de deux zones de radicaux distinctes correspondant à deux règles de réalisation distinctes (une produisant *balay-* et l'autre *balai-*) et d'une unique zone de réalisation du niveau d'exponence, z_{v1}^{exp}, comme indiqué à la figure 2.3. Par exemple, construire la forme *balayent* pour 3PL.PRS.IND met en jeu la règle de réalisation radicale de la zone z_{ay}^{s}, qui produit le radical *balay-*, puis la règle de réalisation d'exponence de la zone z_{v1}^{exp} qui suffixe *-ent*. La construction de la forme alternative *balaient* s'appuie sur la règle de réalisation de l'autre zone radicale, z_{ai}^{s}, qui produit le radical *balai-*, puis la même règle de réalisation issue de la zone z_{v1}^{exp} que pour *balayent*. Pour sa part, *balayons* est obtenu par la règle de la zone radicale de la zone z_{ay}^{s} suivie de la règle de la zone d'exponence z_{v1}^{exp} qui le suffixe par *-ons* pour exprimer 1PL.PRS.IND. Une forme comme **balaions* n'est pas produite pour cette même case puisque l'espace partitionnant de la zone radicale z_{ai}^{s} ne contient pas la structure de traits 1PL.PRS.IND.

Une combinaison valide entre zones de différents niveaux est appelé un *sous-schème flexionnel* [26]. La liste des sous-schèmes correspondant à un lemme donné constitue son *schème flexionnel* I-PAT (cf. figure 2.3 pour le lemme BALAYER).

2.2.1.5 Radicaux supplétifs, formes supplétives

Les entrées lexicales peuvent également spécifier des radicaux supplétifs et des formes supplétives. Dans le cas canonique, la liste S-STEM des radicaux supplétifs et la liste S-FORM des formes supplétives est vide, comme c'est le cas pour BALAYER. Mais dans le cas d'un verbe comme ALLER, chacun des radicaux supplétifs peut être spécifié et associé à un

25. La différence entre niveaux thématiques et niveaux d'exponence est la suivante est liée à la différence entre ce que d'autres modèles, comme la tradition grammaticale latine, appellent la racine et le radical. Prenons un exemple du latin. Une forme comme *accipiō* dispose d'un radical *accip-*, d'une voyelle thématique *-i-* et d'un exposant -ō. Cela correspond à ce que la tradition appelle la racine *accip-* et le radical *accipi-*.

26. Cette définition est incomplète, mais sera complétée sous peu. Elle suffit à ce stade pour comprendre le mécanisme en jeu.

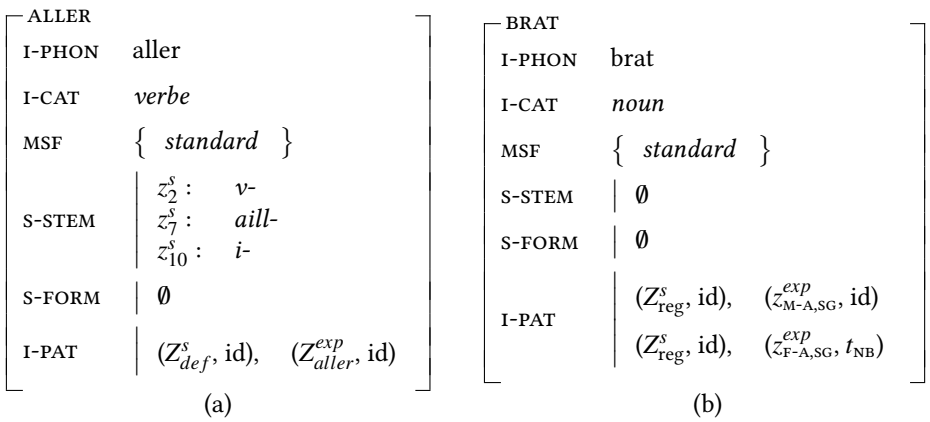


FIGURE 2.4 – Entrées lexicales PARSL pour le verbe français ALLER et le nom serbo-croate BRAT 'frère'. Les indices numériqus associés aux différents radicaux du verbe ALLER_{R} eflètent l'analyse de Bonami et Boyé (2003) du système verbal du français, sur lequel nous reviendrons à la section 4.1.1.

indice radical dénotant l'espace partitionnant de la zone dans laquelle il doit être utilisé en lieu et place de ce que les règles réalisationnelles auraient produit. Les trois radicaux supplétifs *v-*, *aill-* et *ir-* du verbe ALLER sont listés sous S-STEM à la figure 2.4(a). Les formes supplétives sont listés sous S-FORM, associées aux structures de traits morphosyntaxiques qu'elles expriment.

2.2.1.6 Couples réalisationnels et règles de transfert

Pour chaque entrée lexicale, le schème flexionnel I-PAT spécifie un certain nombre de sous-schèmes, qui, comme indiqué plus haut, indiquent chacun une combinaison valide de zones de différents niveaux. Mais en réalité, la situation est légèrement plus complexe, dès lors qu'intervient le phénomène non-canonique de *décalage* (cf. plus haut). En effet, il arrive qu'une zone flexionnelle d'un niveau donné ne soit pas utilisée par un lemme pour remplir les cases qui constituent l'espace partionnnant de la zone. Par exemple, le lemme serbo-croate BRAT 'frère' construit ses formes du pluriel au moyen de règles de réalisation généralement utilisées pour construire des formes au singulier. Dans un tel cas, la zone concernée doit être combinée avec une *fonction de transfert* qui prend en entrée l'une des structures de traits morphosyntaxiques appartenant à l'espace partitionnant de la zone et donne en sortie la structure de traits morphosyntaxique effectivement exprimée par les formes du lemme en question utilisant cette zone. Dans le cas de BRAT, la zone utilisée pour former le pluriel étant une zone exprimant en principe du singulier, la fonction de transfert associée dans le sous-schème correspondant de BRAT est une fonction qui prend en entrée une structure de traits morphosyntaxiques indiquant un nombre singulier et produit en sortie la structure obtenue en remplaçant ce nombre singulier par

le nombre pluriel (cf. figure 2.4). Naturellement, le cas canonique est celui d'une fonction de transfert qui est la fonction identité, notée id. Dans les autres cas, on est en présence d'un phénomène de décalage. Ainsi, un sous-schème est en réalité un ensemble de couples (zone réalisationnelle, fonction de transfert) relevant chacun d'un niveau réalisationnel différent, tels que le couple (Z^s_{reg}, id) dans l'entrée lexicale de BRAT.

2.2.2 Adapter Alexina à PARSL

La première version d'Alexina$_{\text{PARSL}}$ a été développée et utilisée dans le cadre d'un travail sur la flexion verbale du français (Sagot et Walther, 2011 ; Walther et Sagot, 2011a ; cf. section 4.1.1). Cette version reposait sur une version antérieure de PARSL, qui a significativement évolué depuis, conduisant à une refonte d'Alexina$_{\text{PARSL}}$ présentée notamment dans (Sagot et Walther, 2013) [27]. De plus, ont été rajoutés à cette occasion dans Alexina$_{\text{PARSL}}$ plusieurs aspects qui complètent PARSL à la fois sur des aspects que ce dernier n'a pas vocation à couvrir [28] mais aussi pour fournir des moyens de factoriser au mieux les grammaires morphologiques. C'est cette dernière version, dont nous décrivons les grandes lignes à l'annexe B, qui a servi de base à différents travaux, et notamment sur l'allemand (Sagot, 2014 ; cf. section 3.1.3), le latin (italique, indo-européen ; Walther, 2013b), le maltais (sémitique, afro-asiatique, Malte ; Camilleri et Walther, 2012 ; Walther, 2013b) et le khaling (kiranti, sino-tibétain, Népal ; Walther *et al.*, 2013, 2014b ; cf. section 4.1.2).

2.3 En conclusion

Il est important de prendre PARSL et son implémentation Alexina$_{\text{PARSL}}$ pour ce qu'ils sont : un moyen riche et motivé linguistiquement de construire des modèles cohérents de systèmes morphologiques. Plus précisément, PARSL est une proposition de modélisation de la structure des entrées lexicales et des grammaires morphologiques, avec un point de vue constructif. Alexina$_{\text{PARSL}}$, en tant que formalisme d'implémentation destiné à encoder des modèles permettant de produire ou d'analyser effectivement des formes, comporte, contrairement à PARSL, des dispositifs permettant d'écrire des règles de réalisation.

27. La première version de PARSL ne prévoyait que deux niveaux réalisationnels, l'un pour les radicaux et l'autre pour l'exponence. Elle ne permettait pas de modéliser l'ensemble des phénomènes non-canoniques attestés. Mais elle s'est avérée suffisante pour l'étude sur la flexion verbale du français mentionné ci-dessus. De plus, nous n'avions implémenté à l'époque que ce qui, dans PARSL, était nécessaire pour cette étude.

28. Par exemple, PARSL est un modèle de ce que l'on pourrait appeler la macromorphologie flexionnelle, c'est-à-dire l'organisation de la grammaire morphologique et du lexique associé. Mais PARSL n'a pas pour ambition de modéliser ce qu'est une règle flexionnelle, ni même de définir des notions comme celle de bloc de règles telle que définie par Stump (2006) dans en Paradigm Function Morphology.

Alexina$_{\text{PARSLI}}$ répond ainsi à un besoin réel en morphologie descriptive, formelle et computationnelle : représenter d'une façon cohérente les informations permettant de produire les formes de lemmes connus et d'analyser des formes inconnues. Il permet de plus la mise en œuvre de différents types de mesures quantitatives, et notamment de mesures de complexité, comme nous le verrons au chapitre 4.

Mais il faut garder en tête les limitations inhérentes à un tel formalisme. PARSLI ou Alexina$_{\text{PARSLI}}$ ne sont pas des *théories* de la morphologie : ils permettent de *représenter* des systèmes morphologiques de multiples façons. Certes, nous venons d'indiquer et nous proposerons au chapitre 4, comme nous venons de l'indiquer, un moyen quantitatif de comparer différentes représentations Alexina$_{\text{PARSLI}}$ d'un même système morphologique, à l'aide d'une mesure de leur compacité qui repose sur des notions issues de la théorie de l'information. Mais PARSLI ou Alexina$_{\text{PARSLI}}$ ne disent rien sur la pertinence de telle ou telle représentation. Tout d'abord, permettre d'encoder ces informations de façon cohérente et compacte ne constitue pas une modélisation d'une partie de la capacité langagière au sens cognitif : rien, ni dans PARSLI ni dans Alexina$_{\text{PARSLI}}$, ne prétend ni n'est en mesure de modéliser les processus cognitifs à l'œuvre dans la gestion cognitive de ce que l'on dénote par le terme de morphologie flexionnelle Walther (2013b). Ensuite, PARSLI ou Alexina$_{\text{PARSLI}}$ sont des dispositifs strictement synchroniques : de nombreuses idiosyncrasies, visibles notamment à travers les règles morphophonologiques et les différents types de supplétion, sont la conséquence de phénomènes diachroniques. Évidemment, un locuteur n'a pas en tête les états antérieurs de sa langue. Mais entre deux analyses également plausibles cognitivement (quoi que cela veuille dire, et PARSLI n'a rien à dire à ce sujet), on pourrait préférer choisir celle qui corresponde mieux au processus diachronique qui a conduit au système synchronique que l'on décrit. Enfin, PARSLI et Alexina$_{\text{PARSLI}}$, comme tous les modèles de la morphologie, qu'ils soient constructifs ou abstractifs au sens de Blevins (2006), reposent sur des principes sous-jacents qui sont loin d'être faciles à motiver. Autrement dit, PARSLI et Alexina$_{\text{PARSLI}}$ sont des outils de modélisation pertinents pour la représentation de systèmes morphologiques synchroniques, dans une perspective formelle et computationnelle et non cognitive, et donc ni théorique ni explicative. Du reste, si un tel effort de théorisation aurait certainement un impact important, la nature de ce que devrait être une théorie de la morphologie, à valeur explicative, est difficile à cerner. Contraindre une famille de modèles puis prétendre expliquer pourquoi tel type de systèmes morphologiques n'est pas attesté à partir de ces contraintes n'explique que les propriétés de la famille de modèles considérés, et ne constitue pas une théorie de la morphologie.

Chapitre 3

Acquisition automatique d'informations morphologiques

Sommaire

Le développement de ressources lexicales de niveau morphologique est une problématique importante pour le traitement automatique des langues. Nous étudierons aux chapitres 8 à 7 l'impact de telles ressources sur les performances de différents types d'outils d'analyse automatique, et nous concentrons dans ce chapitre sur les techniques de

construction et d'extension de lexiques morphologiques. Ces techniques reflètent la différence structurelle entre les niveaux flexionnel et dérivationnel. Nous nous pencherons donc successivement, dans ce chapitre, sur le développement de lexiques flexionnels, sur celui de lexiques dérivationnels, et enfin sur la prise en compte dynamique des mots inconnus dans des corpus eux-mêmes dynamiques, en l'espèce des dépêches d'agence de presse. Dans ce dernier cas, si l'objectif est l'extension de lexiques flexionnels, les relations dérivationnelles jouent un rôle important.

3.1 Développement de lexiques flexionnels

Les informations lexicales dont on dispose au départ pour le développement d'un lexique flexionnel peuvent varier. Nous évoquerons plusieurs directions de recherche que nous avons suivies et qui ont toutes pour objectif la construction de lexiques morphologiques à partir de données existantes ou construites manuellement pour l'occasion. Elles ont en commun de chercher à construire des entrées morphologiques intensionnelles, et, dans certains cas une grammaire morphologique (sauf si cette dernière fait partie des données d'entrée). Les jeux de données de départ que nous traiterons seront successivement :

— une grammaire morphologique et un corpus brut,

— un lexique extensionnel (sans grammaire morphologique),

— une grammaire morphologique et des données lexicales bruitées et non structurées,

— un lexique extensionnel pour une langue étymologiquement proche.

Comme nous le verrons, les approches suivies dans chacun de ces cas, quoique différentes de par la nature même de la tâche, reposent sur un principe commun : elles cherchent à extraire des données disponibles des régularités significatives quant au système lexico-morphologique de la langue étudiée, régularités destinées dans notre cas à être modélisées en Alexina ou en Alexina PARSLI.

3.1.1 Acquisition automatique de lexiques flexionnels à partir d'une grammaire morphologique et d'un corpus brut

Sauf à développer un lexique morphologique à partir de rien, la situation la moins favorable lorsque l'on développe une telle ressource est de n'avoir à sa disposition qu'un corpus textuel brut. De nombreuses approches ont été proposées pour le développement automatique ou semi-automatique de lexiques morphologiques, notamment pour les langues moins dotées. Certaines de ces approches ne reposent sur aucune connaissance linguistique *a priori*, et relèvent par conséquent du paradigme de l'apprentissage non supervisé de la morphologie (Goldsmith, 2001 ; Creutz et Lagus, 2005 ; Monson *et al.*,

2008). Dans ce telles approches, seul un corpus brut est donné en entrée au système, lequel produit automatiquement soit une segmentation en morphes de tous les mots du corpus soit même un ensemble complet de paradigmes flexionnels, chacun étant associés à un ensemble de lemmes (Snover et Brent, 2001 ; Monson *et al.*, 2008). Toutefois, compte tenu de la complexité et de la richesse des descriptions morphologiques disponibles (formalisées ou non) pour un grand nombre de langues du monde, on peut comprendre le point de vue d'auteurs comme Forsberg *et al.* (2006) selon lesquels il est contre-productif, à la fois en termes de temps de développement et de qualité des résultats, de chercher à reconstituer automatiquement toute cette complexité plutôt que de formaliser les descriptions morphologiques existantes.

Un tel positionnement a conduit plusieurs auteurs, dont nous-mêmes, à proposer des algorithmes d'extraction automatique de lexiques morphologiques à partir de corpus bruts et d'une description formalisée de la morphologie de la langue étudiée. La méthodologie générale est souvent la suivante : on utilise cette description pour construire, à partir des formes inconnues d'un corpus, un certain nombre de lemmes hypothétiques ; diverses techniques permettent alors de faire le tri dans ces lemmes hypothétiques, et d'en faire ressortir certains comme plus vraisemblables. Un certain nombre de travaux mettant en œuvre tout ou partie de ces idées ont été publiés, par exemple pour l'acquisition de lexiques morphologiques du grec (Turcato *et al.*, 2000), du russe (Oliver *et al.*, 2003), du croate (Oliver et Tadić, 2004), des verbes français (Clément *et al.*, 2004), de toutes les catégories du français (Forsberg *et al.*, 2006), du slovaque (Sagot, 2005a), du polonais (Sagot, 2007), des noms allemands (Perera et Witte, 2005), de toutes les catégories de l'allemand (Adolphs, 2008) ou de l'italien (Zanchetta et Baroni, 2005).

Ces travaux diffèrent par la solidité du modèle probabiliste sous-jacent, qui permet la classification des lemmes hypothétiques, et par les indices complémentaires utilisés pour le renforcement de certains lemmes : utilisation de connaissances linguistiques décrites manuellement (contraintes sur les radicaux compatibles avec une classe flexionnelle), utilisation de requêtes sur Internet pour y tester l'existence de formes manquantes, utilisation d'informations relevant de la morphologie dérivationnelle (conversions, affixes dérivationnels, etc.), utilisation d'informations provenant d'un étiqueteur morphosyntaxique reposant sur un état antérieur du lexique (éventuellement restreint aux classes fermées telles que les prépositions, déterminants, conjonctions, etc. ; ceci suppose naturellement la disponibilité préalable d'un lexique, et l'on est alors plutôt dans une perspective d'extension de lexique plutôt que de création).

Le modèle décrit dans (Sagot, 2005a) est le suivant :

1. Lemmatisation ambiguë du corpus : on produit tous les lemmes possibles compatibles avec la grammaire morphologique utilisée et dont au moins une des formes est attestée dans le corpus ;

2. Ordonnancement de ces lemmes candidats à l'aide d'un modèle probabiliste qui met en œuvre le principe suivant : un lemme est d'autant plus vraisemblable que de nombreuses formes différentes en sont attestées dans le corpus, en proportions respectives cohérentes avec ce qui est observé pour les autres lemmes de la même catégorie ; de plus, les liens de morphologie dérivationnelle qui peuvent exister entre deux lemmes tend à renforcer l'un si l'autre est très vraisemblable [1] ;

3. Validation manuelle des lemmes les plus vraisemblables au moyen d'une interface de validation dédiée [2] ; ces validations permettent la mise à jour d'une version du lexique utilisée par les itérations suivantes pour ajuster les paramètres du modèle probabiliste.

L'utilisation de cette méthodologie, sous cette forme ou sous une forme simplifiée, est notamment à l'origine de la toute première version du Le*fff* (Clément *et al.*, 2004), qui n'était alors qu'un lexique morphologique du français restreint aux lemmes verbaux, de la première version du lexique SkLex du slovaque (Sagot, 2005a), de l'étape d'extension du lexique PolLex du polonais, étape destinée à diminuer le nombre de mots inconnus de PolLex au sein du Corpus National Polonais (Sagot, 2007), et également du développement de notre lexique SoraLex du kurde sorani (Walther et Sagot, 2010) [3]. Grâce à ces différents travaux, nous avons pu montrer la pertinence de cette approche : la grammaire morphologique étant développée à la main, on peut en contrôler la qualité et la couverture. Quant aux informations lexicales, le fait qu'elles soient validées manuellement et le caractère itératif du processus de construction de lexique en garantit la précision. La couverture, de son côté, est assurée par l'ancrage en corpus de cette méthode. Enfin, une telle méthode garantit que la ressource construite ne contienne pas d'entrées trop rares, puisque seules des formes attestées dans le corpus de départ peuvent conduire à l'ajout de lemmes dans le lexique. Naturellement, une telle approche dépend fortement du corpus utilisé : les candidats lemmes les mieux classés sont les plus fréquents du ou des domaines concernés, et les lemmes absents du corpus ne seront pas extraits. La construction d'un lexique général requiert donc l'utilisation d'un corpus général, mais à l'inverse il est possible d'utiliser cette technique pour étendre un lexique existant à un

1. Le modèle probabiliste utilisé à cette étape est détaillé dans (Sagot, 2005a).

2. En effet, l'obtention d'un lexique morphologique de qualité satisfaisante nécessite des étapes de validation manuelle des lemmes candidats proposés, typiquement de ceux de plausibilité maximale. Nous avons donc développé une interface web de validation de telles lexiques, dont différentes versions successives ont été utilisées non seulement pour le développement de la première version du Le*fff* (Clément *et al.*, 2004), celui de SkLex (Sagot, 2005a) et l'extension de PolLex (Sagot, 2007), mais également pour identifier des erreurs dans le lexique PerLex du persan, développé par d'autres moyens (Sagot et Walther, 2010b ; Sagot *et al.*, 2011c).

3. Dans tous ces travaux, les grammaires morphologiques utilisées ont été développées dans le formalisme Alexina d'origine — l'utilisation de grammaires Alexina PARSLI nécessiterait de savoir produire un lemmatiseur ambigu à partir de telles grammaires, travail que nous n'avons pas encore réalisé et qui sera plus délicat qu'avec le formalisme morphologique Alexina d'origine.

domaine nouveau à partir seulement d'un corpus brut de ce domaine. Nous présenterons toutefois à la section 3.2 des techniques plus spécifiquement adaptées à une telle tâche.

3.1.2 Construction ou extension d'un lexique flexionnel à partir d'un lexique extensionnel

Pour un certain nombre de langues, des lexiques morphologiques extensionnels ont été développés, qui ont été rendus librement disponibles. Contrairement à un lexique extrait d'un corpus arboré, de tels lexiques ont la propriété de contenir, pour chaque lemme qui y est présent, l'intégralité de son paradigme. Il est donc envisageable de reconstruire, à partir d'un lexique extensionnel, une grammaire morphologique (flexionnelle) et un lexique intensionnel associé qui produisent exactement le même ensemble de formes fléchies. La situation est donc ici fort différente de celle envisagée précédemment. Quand bien même la grammaire ainsi construite est très imparfaite, elle peut servir de point de départ pour le développement d'une grammaire linguistiquement plus motivée (cf. section suivante), ou simplement permettre l'obtention d'un lexique Alexina utilisable directement dans des outils reposant sur de tels lexiques. À plus long terme, on peut envisager la construction automatique d'une grammaire minimisant des mesures de complexité ou de compacité, permettant ainsi la confrontation d'hypothèses concurrentes quant à la structure du système morphologique à l'œuvre — concurrence entre hypothèses produites automatiquement ou avec des analyses produites manuellement.

Nous avons tout d'abord développé une technique simple pour construire un lexique intensionnel Alexina (y compris la grammaire morphologique), destinée exclusivement aux deux premières applications évoquées ci-dessus. Cette technique [4] repose sur l'hypothèse que l'on est en présence d'une morphologie affixale, et construit autant de classes flexionnelles différentes que nécessaires pour que toutes les formes fléchies de chaque lemme puissent être construites au moyen de préfixes et de suffixes accolés à un « radical » commun à toutes les formes du paradigme correspondant. L'hypothèse est faite que la forme de citation est construite sans préfixe, et le fait de se limiter à un radical unique est une limitation forte, notamment dans les cas d'allomorphie ou de supplétion radicale. La pertinence linguistique des grammaires morphologiques ainsi obtenues est donc faible, et ce d'autant plus lorsque l'on est face à une morphologie non strictement concaténative ou que les phénomènes morphophonologiques sont massifs. Mais il existe plusieurs contextes dans lesquels cette approche est suffisante. Un premier exemple serait l'induction de paradigmes complets pour des néologismes ou des formes mal attestées, dès lors que l'on parvient à trouver la bonne classe flexionnelle et la bonne forme de citation pour ces néologismes. Un autre exemple, sur lequel nous reviendrons ci-dessous, est celui de la fusion de lexiques existants.

4. On pourra se reporter à la section 4 de (Sagot, 2014) pour un aperçu plus détaillé.

Nous avons notamment utilisé cette approche sur le lexique Morph-it ! de l'italien (Zanchetta et Baroni, 2005), le lexique SALDO du suédois (Borin *et al.*, 2008) et le lexique MULTEXT de l'espagnol (Ide et Véronis, 1994), afin d'obtenir des lexiques Alexina pour ces langues — dans le cas de l'espagnol, le résultat a servi de première étape pour le développement du lexique Le*ff*e (Nicolas *et al.*, 2008a ; Molinero *et al.*, 2009a,b). L'une des étapes du développement du lexique DeLex de l'allemand en a également fait usage (Sagot, 2014) [5].

Naturellement, les classes flexionnelles ainsi construites rassemblent des unités lexicales dont le comportement flexionnel est identique. Si l'on dispose déjà d'un lexique flexionnel (intensionnel) L pour la langue sur laquelle on travaille, on peut s'appuyer sur les entrées lexicales en commun entre L et un autre lexique extensionnel L' pour procéder comme suit : (1) construction d'une version intensionnelle de L' avec la méthode décrite ci-dessus ; (2) construction automatique d'une table de correspondance entre classes flexionnelles acquises automatiquement pour L' et classes flexionnelles de la ressource d'origine L ; (3) grâce à cette table de correspondance, production d'une version intensionnelle de L' qui s'appuie sur la même grammaire morphologique que L. C'est ainsi que nous avons converti en des lexiques « compatibles-Le*ffff* » les principaux lexiques du français, et notamment Morphalou, le DELA, le Lexique des Verbes Français, ProLex et le Wiktionnaire — dans ce dernier cas, après une étape non triviale d'extraction dont une version étendue a été utilisée pour le développement du lexique DeLex (cf. section 3.1.3). À chaque fois, les contraintes associées par la grammaire morphologique du Le*ffff* à chaque classe permet de détecter automatiquement des erreurs dans les ressources d'origine.

Convertir ainsi un lexique flexionnel L' de sorte qu'il utilise les mêmes classes flexionnelles qu'un lexique L permet la fusion de ces deux lexiques. C'est ainsi, par exemple, que nous avons construit le niveau morphologique du lexique Alexina Le*ff*e de l'espagnol, en convertissant le lexique MULTEXT de cette langue au format Alexina, puis en le complétant par fusion avec le lexique USC (Álvarez *et al.*, 1998), le lexique ADESSE (García-Miguel et Albertuz, 2005) et le lexique de la Spanish Resource Grammar (Marimon *et al.*, 2007) — pour plus de détails, on pourra se reporter à (Nicolas *et al.*, 2008a ; Molinero *et al.*, 2009a,b). De même, une première version de notre lexique Alexina du persan, PerLex (Sagot et Walther, 2010b,a), construite à partir du corpus BijanKhan (BijanKhan, 2004) et d'autres sources d'informations lexicales [6], a été améliorée grâce, entre autres, à une fusion avec le lexique du persan développé dans le cadre du projet

5. Nous avons également entamé le développement d'une approche bien plus riche pour réaliser cette même tâche, approche prévue notamment pour détecter et encoder en Alexina les alternances de radicaux et, à terme, induire des règles morphophonologiques permettant de minimiser le nombre de règles de réalisation. Cette approche, en cours de développement et d'application à la flexion verbale du grec moderne, n'est pas encore suffisamment aboutie pour que nous en fassions mention plus avant.

6. Nous ne décrirons pas ici le processus mis en œuvre à cet égard. On pourra se reporter aux publications citées pour plus d'informations.

MULTEXT-East (QasemiZadeh et Rahimi, 2006 ; Erjavec, 2010), aboutissant ainsi, après validation manuelle partielle, à la deuxième version de PerLex (Sagot *et al.*, 2011b,c) [7]. De tels travaux permettent d'exploiter au mieux des ressources multiples existant pour une même langue, malgré les divergences entre elles quant à la façon de modéliser la morphologie flexionnelle de la langue considérée.

3.1.3 Acquisition automatique de lexiques flexionnels à partir d'une grammaire morphologique et de données lexicales bruitées et non structurées

Depuis l'essor des ressources collaboratives modifiées et complétées librement par la communauté des internautes, au premier rang desquelles les ressources de type wiki (les encyclopédies wikipedia et les dictionnaires wiktionary, et notamment le Wiktionnaire, pour le français), une nouvelle source d'informations lexicales est librement disponible, dont la couverture est parfois bien plus importante que les lexiques extensionnels existants. Les ressources de type wiki ne sont toutefois pas formalisées sous la forme de lexiques intensionnels et de grammaires morphologiques tels que nous les avons définis au chapitre précédent. Bien qu'il soit possible de télécharger l'intégralité de ces ressources, il est donc nécessaire de fournir un travail important pour exploiter les informations qu'elles contiennent et les valoriser sous la forme d'un lexique morphologique [8]. Le travail requis est d'autant plus important que ces ressources sont bruitées, tant sur la forme (la syntaxe des articles composant ces ressources n'étant ni triviale ni vérifiée automatiquement) que sur le fond (les informations qui y sont présentes ne sont pas toutes valides) [9].

Nous avons mis en œuvre cette approche pour développer DeLex, un nouveau lexique flexionnel de l'allemand, à partir du Wiktionary allemand (Sagot, 2014). En effet, et de façon surprenante, il semble n'avoir existé avant 2014 aucun lexique morphologique

7. Une partie des entrées nominales et adjectivales du Le*fff* est ainsi issue du lexique MULTEXT du français (Veronis, 1998), mais le Le*fff* a ensuite été complété principalement par ajouts manuels, parfois guidé par les résultats de son utilisation dans l'analyse syntaxique de gros corpus (cf. chapitres 5 et 9). En revanche, sauf dans des cas très spécifiques (adverbes en *-ment*, verbes en *-iser* et *ifier*), le Le*fff* n'a pas été complété par la suite par fusion avec d'autres ressources lexico-morphologiques. La raison en est la couverture alors déjà suffisante du Le*fff*, et, à l'inverse, les nombreuses entrées rares, archaïques ou erronnées présentes dans les autres ressources, comme discuté précédemment. En revanche, ces autres ressources, une fois converties au format Alexina, ont été mises à contribution dans plusieurs travaux présentés dans ce chapitre, notamment pour le développement d'un lexique dérivationnel et pour la construction d'une chaîne d'analyse des mots inconnus (section 3.3.1) et de construction de nouvelles entrées lexicales candidates à partir de flux textuels (section 3.2).

8. Pour le français, nous avons évoqué ci-dessus le lexique morphologique GLÀFF extrait du Wiktionnaire, qui contient également des représentations phonétiques (ou plutôt phonémiques) de la plupart des formes fléchies (Sajous *et al.*, 2013 ; Hathout *et al.*, 2014).

9. Un autre ensemble de ressources lexicales non structurées est disponible, à savoir les lexiques développés par des linguistes de terrain dans le cadre de leurs travaux descriptifs. Lorsqu'ils sont disponibles et exploitables informatiquement, ces lexiques prennent souvent la forme de dictionnaires électroniques à partir desquels l'extraction d'informations lexicales formalisées n'est pas non plus triviale, bien qu'ils soient naturellement pas (ou peu) bruités. C'est ainsi que nous avons développé le lexique Alexina du kurde kurmanji, développé notamment à partir des travaux descriptifs de (Thackston, 2006).

de l'allemand librement disponible dont la couverture fût importante, comme relevé du reste par Adolphs (2008). Le seul lexique disponible était, semble-t-il, le lexique à couverture moyenne distribué avec l'analyseur morphologique `morphisto` (Zielinski et Simon, 2009) qui repose sur SMOR, modélisation de la morphologie de l'allemand sous la forme d'un d'automate fini (Schmid *et al.*, 2004) [10]. `Morphisto` contient 18 624 entrées parmi lesquelles 17 749 « radicaux de base », lesquels sont, en première approximation, principalement des lemmes mais aussi des radicaux et des formes supplétives [11]. Indépendamment du développement de DeLex, une autre équipe extrait un lexique flexionnel à partir du Wiktionary allemand, nommé Zmorge (Sennrich et Kunz, 2014), mais en mettant moins fortement l'accent sur le développement d'une grammaire morphologique linguistiquement pertinente, ce lexique couplé à la grammaire morphologique SMOR.

Outre l'intérêt qu'il y avait à combler un manque dans la communauté, un autre intérêt de développer DeLex résidait dans les caractéristiques de la morphologie de l'allemand. En effet, la morphologie flexionnelle allemande n'est pas strictement concaténative. Elle implique par ailleurs de nombreux types de phénomènes non canoniques (cf. le chapitre précédent, et pour référence, Corbett, 2003). Trois d'entre eux sont particulièrement répandus : l'alternance de radicaux, le syncrétisme et la surabondance au sens de Thornton (2011). Plus précisément, les radicaux supplétifs et les alternances vocaliques sont particulièrement répandues dans la flexion verbale, la surabondance est généralisée, notamment par le caractère facultatif de la voyelle *e* dans de nombreuses formes nominales et en finale des impératifs présents singuliers, et l'utilisation de différents niveaux et blocs (cf. chapitre précédent) est une façon naturelle de formaliser le système morphologique dans son ensemble [12].

Nous renvoyons à (Sagot, 2014) pour un bref aperçu des propriétés morphologiques de l'allemand, lequel explique plus en détails à quelle point la morphologie flexionnelle de cette langue est un banc d'essai pertinent pour mettre en œuvre et valider le formalisme Alexina décrit au chapitre précédent, et par ce biais le modèle de la morphologie flexionnelle sur lequel il repose. L'idée générale de la méthodologie proposée, qui est détaillée dans (Sagot, 2014) :

10. `Morphisto` est distribué sous licence Creative Commons 3.0 BY-SA, mais SMOR l'est sous licence GPL v2, ce qui n'en permet pas l'utilisation commerciale.

11. Par exemple, le dictionnaire librement distribué avec la plateforme Unitex (Paumier, 2003) contient 300 000 formes fléchies qui représentent 10% du lexique non-libre CISLEX (Langer *et al.*, 1996).

12. Par exemple pour la modélisation de la flexion adjectivale (comparatifs et superlatifs étant des formes fléchies de l'adjectif, les marques de genre/nombre/cas étant les mêmes que pour les formes de base).

1. Extraction des paradigmes partiels fournis par le wiktionary [13] et construction d'un lexique intensionnel partiel à partir de ces paradigmes partiels grâce à la méthode évoquée à la section précédente [14] ;

2. Développement manuel d'une description morphologique complète pour la langue considérée ;

3. Conversion du lexique intensionnel partiel en lexique intensionnel faisant usage des schèmes flexionnels décrits dans la grammaire développée manuellement, selon la méthode indiquée à la section précédente.

On peut noter que ce processus permet à la fois de produire un lexique satisfaisant du point de vue morphologique et de couverture importante, mais également, comme mentionné à la section précédente, d'identifier de nombreuses erreurs présentes dans le wiktionary [15].

L'application de cette méthodologie au développement de DeLex, après création des lexiques des autres classes fermées grâce au corpus TIGER Brants *et al.* (2002) ; Smith (2003) et par un travail manuel, a permis la construction d'un lexique de 63 017 entrées intensionnelles, dont 6 530 lemmes adjectivaux, 39 670 lemmes nominaux, 4 899 lemmes verbaux et 904 lemmes adverbiaux [16]. Une fois fléchies, ces entrées intensionnelles produisent plus de 2,3 millions d'entrées morphologiques extensionnelles [17].

Nous avons réalisé plusieurs évaluations différentes du lexique DeLex, pour lesquelles nous renvoyons à (Sagot, 2014), et dont nous résumons ici les conclusions. Tout d'abord, DeLex couvre 93,1% des mots du corpus TIGER, hors entités nommées. Par ailleurs, la couverture de DeLex est bien supérieure à celle de `morphisto`, malgré l'existence de nombreuses entrées lexicales présentes (sous une forme ou une autre) dans `morphisto`

13. Dans certains cas, notamment pour l'allemand, ces paradigmes partiels contiennent des cases surabondantes.

14. Ce lexique est partiel dans la mesure où les paradigmes fournis par le wiktionary ne sont pas complets

15. C'est le cas, par exemple, lorsque la forme de citation d'une entrée lexicale n'est pas l'une des formes fléchies produites par notre lexique alors qu'elle le devrait — cela dépend de la façon dont le lexique est représenté — (6 128 lemmes éliminés sur 55 574, pour les adjectifs, noms et verbes de DeLex), ou encore la grammaire morphologique de DeLex (développée à la main) ne peut pas fléchir une entrée extraite de wiktionary, après conversion de sa classe flexionnelle obtenue automatiquement en classe flexionnelle de DeLex, notamment en raison d'une contrainte sur les radicaux admissibles (cf. chapitre précédent). Ce cas s'est révélé rare concernant le Wiktionary allemand et DeLex, mais massif lors de la conversion du wiktionnaire français vers les classes du L*efff* (7 324 sur 77 683). Parmi ces derniers cas, on trouve aussi naturellement des erreurs liées au processus de conversion (notamment lorsque le paradigme du lemme n'est pas donné, et se réduit donc par erreur à un lemme invariable, souvent mis en correspondance avec la classe flexionnelle la plus fréquente, éventuellement non compatible).

16. On notera qu'en allemand les adjectifs sont souvent utilisés comme adverbes, les lemmes adverbiaux sont donc les adverbes non adjectivaux comme SCHON 'déjà', ZUGLEICH 'en même temps', ETWA 'à peu près', etc.).

17. Plus précisément, sont produites environ 441 000 formes adjectivales, 301 000 formes nominales, 486 000 formes verbales et 908 formes adverbiales. DeLex est, comme tous les lexiques Alexina, librement disponible sous licence LGPL-LR.

qui sont absentes de DeLex [18]. Enfin, l'utilisation de DeLex en complément à un corpus annoté pour l'entraînement d'un analyseur morphosyntaxique permet d'en améliorer les performances (cf. chapitre 8).

La méthodologie utilisée pour développer DeLex est applicable à toute langue pour laquelle est disponible un wiktionary de taille suffisante et fournissant des paradigmes partiels. Une autre possibilité serait d'extraire directement les patrons flexionnels utilisés pour produire les tables flexionnelles visualisables dans certains wiktionary, mais ces tables ne constituent pas des grammaires morphologiques linguistiquement intéressantes, et il conviendrait là aussi de redévelopper, comme nous l'avons fait pour DeLex, une grammaire morphologique pertinente à partir des données brutes. De façon plus générale, ce type de travaux montre la faisabilité du développement de lexiques morphologiques formalisés à partir de données faiblement structurées et souvent bruitées. Outre les wiktionary, une autre source d'informations lexicales morphologiques de nature comparable est fournie par les grammaires produites par des linguistes de terrain, dès lors qu'elles sont accompagnées de lexiques de taille suffisante. C'est à partir d'une telle ressource (Thackston, 2006) que nous avons développé un lexique Alexina pour le kurde kurmanji, KurLex (Walther *et al.*, 2010). C'est également le cas pour le khaling (kiranti, sino-tibétain, Népal), pour lequel le travail de terrain de Guillaume Jacques (cf. par exemple Jacques *et al.*, 2012) a servi de base au développement du lexique Alexina KhaLex (Walther *et al.*, 2013, 2014b), restreint à ce jour aux entrées verbales.

3.1.4 Acquisition automatique de lexiques flexionnels à partir d'un lexique flexionnel pour une langue étymologiquement proche [19]

Le développement de ressouces et d'outils pour le traitement automatique de langues peu dotées, et notamment de langues régionales apparentées à des langues plus importantes numériquement ou politiquement, pose des difficultés particulières liées au manque de données : le volume de données textuelles numériques (corpus ou lexiques, y compris collaboratifs) est souvent réduit, voire nul, notamment dans le cas de langues ne disposant pas de systèmes d'écriture ; de plus, aucune de ces données textuelles,

18. Pour les noms, un certain nombre de ces lemmes absents de DeLex sont de noms propres (Brasilien 'Brésil', Brazzaville 'id.'), mais également de nombreux mots absents de la version du Wiktionary allemand utilisée. C'est également le cas de la majorité des lemmes absents de DeLex dans les autres catégories. On voit ici à la fois une limite et une qualité de l'approche : elle ne permet pas d'obtenir directement une ressource plus couvrante que le wiktionary utilisé, mais il est relativement facile de mettre à jour la ressource au fur et à mesure que ce wiktionary se complète. Nombreux sont les mots manquants de DeLex mais présents dans la dernière version du wiktionnaire allemand, ce qui montre qu'une telle mise à jour permettrait de compléter efficacement DeLex. Nous n'avons toutefois pas encore procédé à une telle mise à jour, et n'avons donc pas de données quantitatives à ce sujet.

19. Le travail présenté dans cette section a été réalisé en collaboration avec Yves Scherrer, dans le cadre du LabEx EFL (opération LR2.2 de l'Axe 6 « Ressources Linguistiques », axe dont j'ai coordonné la conception et dont j'ai été responsable jusqu'en début 2015, avant d'en devenir responsable adjoint). Ces travaux ont donné lieu à plusieurs publications (Scherrer et Sagot, 2013b,a, 2014).

ou très peu, ont en général été annotées ; enfin, l'orthographe est rarement homogène car rarement standardisée. Si, dans le cas de langues étudiées par des linguistes de terrain, nous avons vu qu'un lexique morphologique est parfois construit comme partie prenante de l'effort de description, ce n'est pas toujours le cas pour des langues régionales ou minoritaires. Et quand bien même elles existent, de telles ressources lexicales ne sont pas toujours disponibles librement. Malgré ces obstacles, qui rendent difficilement applicables les techniques évoquées jusqu'ici, et dès lors que des données textuelles sont effectivement disponibles, la proximité d'avec des langues mieux dotées permet également d'envisager des approches reposant sur les ressources annotées disponibles pour ces dernières. Toutefois, les ressources qui seraient le plus directement exploitables, à savoir des lexiques bilingues et les corpus alignés, n'existent souvent pas.

Nous avons ainsi développé et mis en œuvre une chaîne générique permettant de transférer des annotations morphosyntaxiques disponibles dans une langue donnée, que nous appellerons la *langue dotée (LD)*, vers une *langue non dotée (LND)* apparentée, étymologiquement proche, sans utiliser de ressources bilingues telles que des lexiques bilingues ou des corpus parallèles [20]. On suppose donc disponibles les ressources suivantes : un corpus textuel brut pour la LND, un corpus textuel brut pour la LD et un lexique morphosyntaxique, c'est-à-dire un inventaire de couples (forme, catégorie), pour la LD.

Cette approche repose sur deux hypothèses. D'une part, au niveau lexical, les deux langues partagent un grand nombre de cognats, c'est-à-dire de couples de mots formellement similaires qui sont en relation de traduction. D'autre part, au niveau structural, on suppose que l'ordre des mots est similaire dans les deux langues, et qu'un inventaire d'étiquettes morphosyntaxiques unique peut être utilisé pour les deux langues. Ces hypothèses permettent de chercher à transférer l'étiquette morphosyntaxique d'un mot de la LD vers son équivalent de traduction en LND.

Notre approche, décrite notamment dans (Scherrer et Sagot, 2014), se décompose en deux phases principales, détaillées à la figure 3.1. La première vise à induire un lexique bilingue (un inventaire de couples de la forme $\langle w_{\mathrm{LND}}, w_{\mathrm{LD}} \rangle$) à partir de corpus monolingues

20. Cette tâche est donc très différente de ce que l'on appelle en général l'*étiquetage morphosyntaxique non supervisé*. En effet, on trouve dans la littérature, depuis Merialdo (1994), de nombreux articles qui définissent sous ce nom la tâche consistant à effectuer un étiquetage morphosyntaxique sans corpus d'entraînement mais grâce à un lexique morphosyntaxique préexistant. Autrement dit, on projette un lexique sur un corpus brut, et la tâche revient alors à désambiguïser l'annotation ainsi obtenue. Dans la plupart des cas, des techniques d'apprentissage automatique sont mises en place afin d'induire un modèle probabiliste (cf. cependant Brill, 1995). Le modèle le plus populaire pour cette tâche est celui des Modèles de Markov Cachés (HMM) (Merialdo, 1994 ; Goldwater et Griffiths, 2007 ; Goldberg et Tsarfaty, 2008 ; Ravi et Knight, 2009), mais d'autres modèles ont également été proposés, et notamment des modèles discriminants (Smith et Eisner, 2005). Différentes techniques d'apprentissage ont été utilisées, et des connaissances linguistiques (en plus du lexique) sont parfois intégrées, notamment pour initialiser les paramètres du modèle HMM. À l'inverse de ces travaux, nous ne présupposons ici la disponibilité d'aucune ressource pour la LND, mis à part des données textuelles brutes.

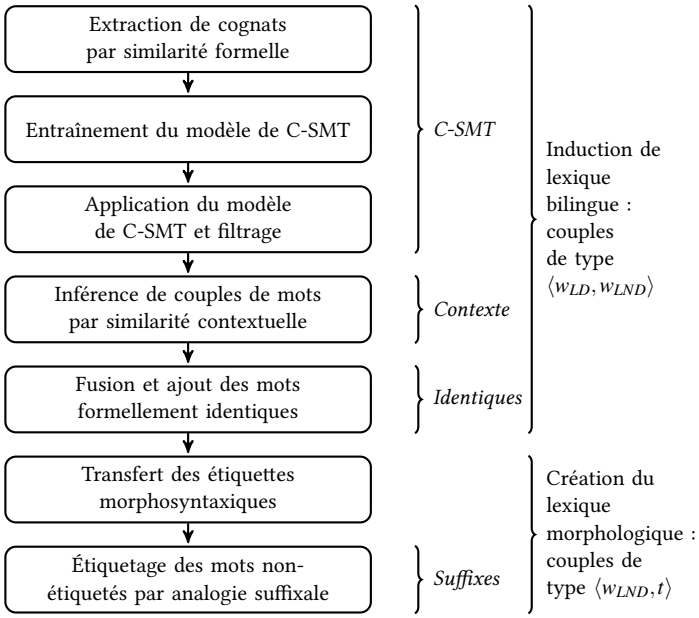


FIGURE 3.1 – Aperçu de notre algorithme de transfert lexico-morphologique entre langues proches. Dans chacune des deux grandes phases sont identifiées les étapes contribuant à construire le lexique pour la LND (cf. table 3.2).

dans les deux langues, grâce à la détection automatique de cognats puis au moyen d'une méthode de similarité contextuelle parallèle. La seconde consiste à transférer les catégories morphosyntaxiques d'un élément à l'autre des paires ainsi obtenues. L'étiquetage complet du corpus brut en LND est alors possible par analogie suffixale. Nous renvoyons aux publications traitant en détail la méthode décrite ici (Scherrer et Sagot, 2013b,a, 2014) pour un état de l'art détaillé sur les travaux ayant un lien avec l'une ou l'autre de ces étapes, et notamment les travaux sur l'inférence de lexiques bilingues (Koehn et Knight, 2002), plus spécifiquement sur l'extraction de couples de cognats Mann et Yarowsky (2001) et notamment sur la mesure de similarité lexicale BI-SIM (Kondrak et Dorr, 2004 ; Inkpen *et al.*, 2005), sur l'utilisation de techniques de traduction automatique utilisant chaque caractère (et non chaque mot) comme unité élémentaire (*character-level statistical machine translation*, dorénavant C-SMT ; Vilar *et al.*, 2007 ; Tiedemann et Nabende, 2009 ; Karanasou et Lamel, 2011 ; Beinborn *et al.*, 2013), sur la similartié contextuelle inter-langue (Fung, 1998 ; Rapp, 1999 ; Fišer et Ljubešić, 2011 ; Xu *et al.*, 2011) et sur le transfert d'informations morphosytaxiques dans des situations plus favorables, notamment parce que l'on dispose de corpus parallèles alignés (Yarowsky *et al.*, 2001 ; Das et Petrov, 2011 ; Duong *et al.*, 2013) ou d'un analyseur morphologique pour la langue cible (Feldman *et al.*, 2006).

Nous avons appliqué et évalué notre approche sur trois ensembles de langues :

Langue	Corpus brut (Wikipedia)			Corpus annoté pour extraction du lexique de référence		
	#phrases	#tokens	#types	Nom	#types	#étiquettes
aragonais	335 091	5 478 092	215 809		–	
asturien	226 789	3 600 117	201 417		–	
galicien	1 955 291	32 240 505	674 848		–	
catalan 500k	22 876	499 978	41 908		–	
catalan 140M	7 939 544	139 160 258	1 712 078		–	
portugais	12 611 706	197 515 193	2 252 337	CETEMPúblico [22]	107 235	117
espagnol	23 381 287	431 884 456	3 451 532	AnCora-ES [23]	40 148	42
néerlandais	33 361	499 991	52 502		–	
palatin	28 149	318 926	51 038		–	
allemand	42 127 804	612 658 190	8 673 998	TIGER [24]	85 691	55
cachoube	25 620	198 560	40 805		–	
bas-sorabe	28 352	265 580	48 189		–	
haut-sorabe	106 299	891 941	104 319		–	
slovaque	2 555 779	30 114 232	1 091 474		–	
tchèque	6 642 402	85 579 006	1 934 787	PDT 2.5 [25]	55 947	57
polonais	16 639 594	206 372 541	3 264 129	NKJP [26]	132 664	29

Tableau 3.1 – Corpus utilisés pour les expériences de transfert lexico-morphologique entre langues proches

— cinq langues romanes de la péninsule ibérique : l'espagnol et le portugais sont les LD, l'aragonais, l'asturien, le galicien et le catalan jouant le rôle de LND (pour le catalan, qui est une langue dotée, il s'agit de disposer de gros volumes de données de référence, notamment pour étudier l'impact de la taille du corpus brut) ;

— trois langues germaniques occidentales, l'allemand étant la LD alors que le palatin (Pfälzisch) et le néerlandais jouent le rôle de LND (le néerlandais, qui est une LD, permettant ici encore de disposer de données d'évaluation en grande quantité) ;

— cinq langues slaves occidentales, le tchèque et le polonais étant les LD alors que le slovaque, le haut-sorabe, le bas-sorabe et le cachoube sont des LND.

Dans nos expériences, nous extrayons les corpus bruts des wikipedia de la LD et de la LND et le lexique morphosyntaxique pour la LD d'un corpus arboré [21]. Les corpus bruts sont utilisés pour la tâche d'induction de lexiques bilingues, alors que le lexique morphosyntaxique est nécessaire pour la tâche d'étiquetage morphosyntaxique du corpus en LND. La table 3.1 récapitule les données utilisées dans nos expériences.

21. On notera que ce lexique morphosyntaxique pourrait être obtenu par d'autres moyens, sans qu'un corpus annoté morphosyntaxiquement ne soit nécessaire. On pourrait par exemple extraire ce lexique morphosyntaxique à partir d'un lexique morphologique, en faisant l'hypothèse que les catégories (morphologiques) fournies par le lexique morphologique soient assimilables, au moins approximativement, à des catégories morphosyntaxiques.

22. http://www.linguateca.pt/CETEMPublico/

23. http://clic.ub.edu/corpus/ancora

24. http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

25. http://ufal.mff.cuni.cz/pdt2.5/

26. http://nkjp.pl

	C-SMT	Contexte	Identiques	Suffixes	Total	#étiquettes
asturien ← espagnol	85,6%	91,2%	91,2%	49,7%	85,4%	42
catalan ← espagnol 500k	84,0%	95,9%	96,2%	47,7%	85,9%	42
catalan ← espagnol 140M	74,7%	93,6%	99,4%	60,0%	89,1%	42
néerlandais ← allemand	50,9%	81,7%	78,3%	31,7%	59,0%	55
palatin ← allemand	72,9%	81,7%	81,9%	35,9%	65,1%	55
haut-sorabe ← tchèque	67,1%	93,4%	96,3%	77,2%	83,6%	57
slovaque ← tchèque	86,8%	94,7%	97,8%	82,0%	91,6%	57
palatin ← tchèque	68,8%	85,9%	95,9%	67,0%	77,6%	57

Tableau 3.2 – Précision de l'étiquetage des corpus en LND, avec résultats des tokens couverts pour chaque étape (cf. figure 3.1) et résultas globaux. Les en-têtes de lignes sont de la forme LND ← LD.

La précision de l'étiquetage ainsi obtenu a été évaluée pour sept paires de langues à partir de l'annotation manuelle de 30 à 100 phrases des LND concernées. Les résultats, fournis à la table 3.2, montrent que notre approche donne des résultats satisfaisants sauf pour les langues germaniques. Le score le plus élevé, 91,6%, est obtenu sur le slovaque en partant du tchèque comme LD.

L'une des principales limitations de notre approche réside dans l'hypothèse que l'inventaire de catégories et le même pour la langue LND et pour la langue LD [27]. Malgré cela, et bien que l'ambiguïté ne soit pas prise en compte, la précision des étiquetages obtenus est généralement satisfaisante. Elle est en réalité du même ordre de grandeur que ce qu'obtiennent les techniques d'étiquetage non supervisées, techniques qui reposent pourtant sur un lexique associant catégories et mots. Ici, nous ne disposons pas d'une telle ressource pour la langue, mais seulement pour une langue proche, sans qu'un corpus parallèle ne nous permette d'inférence directe. Notre approche permet donc d'initier le développement rapide d'un lexique morphosyntaxique, et même d'un corpus annoté morphosyntaxiquement, pour une langue peu dotée à partir de ressources concernant une langue typologiquement proche.

27. Il ne nous a pas été possible de réaliser une analyse quantitative et qualitative détaillée pour toutes les paires de langues, mais nous pouvons illustrer ce phénomène sur notre corpus de polonais annoté manuellement avec les catégories du corpus tchèque :

Nie chcieliśmy rozwiązywać zespołu […]
'Nous ne voulions pas dissoudre le groupe […].'

Les deux premiers mots, *nie chcieliśmy* 'nous ne voulions pas', sont constitués d'un clitique négatif *nie* et d'une forme verbale finie passée du verbe chieć 'vouloir'. Mais ce clitique négatif n'existe pas en tchèque, et il n'y a pas de forme synthétique correspondant à *chcieliśmy*, bien que des morphes cognats soient utilisés dans les deux langues. En effet, le tchèque et ses conventions orthographiques regroupent le marqueur de négation *ne* et une forme de participe passé — *nechtěli* — puis considère la forme d'auxiliaire *jsme* comme un mot distinct. Il n'y a donc pas de façon satisfaisante d'étiqueter *nie* ou *chcieliśmy* avec les catégories du tchèque. Dans de tels cas, au cours de l'annotation manuelle, nous avons fait usage d'étiquettes absentes du jeu d'étiquette de la LD : toute prédiction faite sur de tels mots par notre système conduira à une erreur.

3.2 Extension dynamique de lexiques flexionnels à partir d'un flux textuel [28]

Tout comme les dictionnaires de langues, inévitablement lacunaires, les lexiques utilisés pour des applications en TAL doivent être régulièrement complétés afin de refléter au plus près les réalités linguistiques et limiter ainsi l'incomplétude lexicale. Cependant ce processus continu de mise à jour ne peut suffire à lui seul, ne serait-ce que par le coût humain d'une telle tâche, et ce malgré l'application de techniques telles que celles décrites jusqu'ici. Il est donc utile de disposer de modules d'analyse permettant d'extraire automatiquement de nouvelles entrées lexicales et les ajouter, après validation manuelle ou automatique, dans des ressources lexicales. La mise au point de tels outils est plus particulièrement intéressante pour le traitement des données textuelles récentes, voire des corpus dynamiques comme un flux de dépêches d'agence, produites en temps quasi-réel. Nous renvoyons à (Sagot *et al.*, 2013) pour une description plus détaillés de nos travaux à ce sujet, qui ont porté uniquement sur le français.

3.2.1 L'incomplétude lexicale et les néologismes

Étant donné un outil de TAL et un texte à traiter, certains tokens sont des *inconnus* : à partir du lexique, l'outil ne parvient pas à les analyser comme mots-formes simples ou combinaisons régulières de tels mots-formes (par exemple, en français, *donne-moi* est inconnu en tant que tel des lexiques de référence mais analysable comme combinaison typographique des mots-formes *donne* et *-moi*). Nous avons utilisé comme référence le Le*fff* et l'ensemble des mentions d'entités nommées répertoriées dans notre base *Aleda* (Sagot et Stern, 2012) [29].

De nombreuses typologies des inconnus ont été proposées, plus ou moins profondes et couvrant plus ou moins de cas [30]. Nous avons choisi de définir une typologie simple et orientée vers les traitements, couvrant tous les types d'inconnus et adaptée de celle que nous avons nous-même proposée dans (Blancafort San José *et al.*, 2010) :

— les **tokens invalides**, induits notamment par des erreurs de tokenisation ;

28. Le travail présenté dans cette section a été réalisé en collaboration avec Damien Nouvel, Virginie Mouilleron et Marion Baranes, et financé principalement par le projet ANR EDyLex, projet ANR dont j'étais le porteur, mais également par l'entreprise viavoo au sein de laquelle travaillait Marion Baranes. Il a été publié dans (Sagot *et al.*, 2013). Ce travail s'appuie sur des études antérieures menés là aussi dans le cadre d'EDyLex, en collaboration avec Helena Blancafort San José et Javier Couto de l'entreprise Syllabs, Denis Teyssou de l'Agence France-Presse, ainsi que Gaëlle Recourcé et Rosa Stern, alors à Alpage (Blancafort San José *et al.*, 2010). Le module d'analyse des néologismes dérivationnels par analogie a été développé dans le cadre de la thèse de Marion Baranes (2015) dont j'étais encadrant principal et co-directeur. Il en est de même pour le module de détection des emprunts non adaptés (Baranes, 2012).

29. Dans ce travail, nous avons laissé de côté les inconnus contextuels, c'est-à-dire les tokens qui ne sont connus de la référence que comme composants de composés mais qui apparaissent dans d'autres contextes (par exemple, *instar* si on le trouvait ailleurs que dans le composé *à l'instar de/du*).

30. Sablayrolles (1997) en recense ainsi une centaine.

- les **inconnus orthographiques**, produits de façon consciente (économie scripturale), par erreur (mauvaise connaissance de l'orthographe), ou en raison d'instabilités orthographiques (notamment pour les emprunts, les constructions préfixales ou les associations : *co-fondateur*, *coproducteur*, *microalgues*, *micro-ondes*, *électro-mécanique*, *électroencéphalogramme*) ;
- les **inconnus typographiques** (absence de tirets ou de blancs typographiques obligatoires) ;
- les **nombres**, **sigles** et autres unités de ce type (*A380*, *L-334-1*) ;
- les **emprunts non adaptés**, qui ne sont pas encore rentrés dans le système morphologique de la langue et ne disposent pas encore de paradigmes morphologiques complets
- les **inconnus lexicaux**, formes correctes absentes des ressources de référence (emprunts adaptés, créations lexicales, entités nommées nouvelles ou rares, mentions inconnues d'entités connues, etc.) ; parmi eux, il convient de distinguer les mentions d'entités nommées d'une part et le reste d'autre part, que nous qualifierons de **néologismes** dans la suite de chapitre [31].

L'une des difficultés de toute approche d'enrichissement de lexiques morphologiques est de faire le départ entre les inconnus lexicaux, seuls susceptibles d'être intégrés à un lexique morphologique, et les autres tokens inconnus (cf. exemple `coproducteurs` vs. `L-334-1`). Si, dans certains cas, il s'agit d'une tâche relativement aisée (sigles, nombres), distinguer un néologisme d'un inconnu orthographique ou d'un emprunt non adapté est moins immédiat. En nous limitant uniquement aux inconnus lexicaux morphologiquement analysables, notre objectif est ici triple : (1) mettre en évidence les phénomènes constructionnels dont procèdent les néologismes, (2) montrer qu'il est possible d'identifier et d'analyser automatiquement ces néologismes, et (3) étendre ainsi automatiquement le lexique morphologique de référence retenu, ici le Le*fff*. [32]

3.2.2 Données : le flux de dépêches AFP

Nous avons travaillé sur des volumes importants de de dépêches en français de l'Agence France-Presse (AFP). L'AFP illustre les besoins que peuvent rencontrer des industriels en termes de technologies de traitement automatique des langues, et la complexité des

31. Nous considérons donc comme étant un néologisme toute unité lexicale valide qui est nouvelle par rapport aux lexiques de référence, et non, comme c'est souvent le cas, par rapport à un usage supposé connu et vérifiable. Puisqu'il ne s'agit pas d'inconnus, nous ne traitons pas non plus des cas où une forme graphique connue est employée avec une catégorie inconnue du lexique (conversion) ou avec un sens nouveau (néologie sémantique).

32. Nous renvoyons à (Sagot *et al.*, 2013) pour un bref panorama des travaux antérieurs sur l'étude et la description de la néologie, en linguistique comme en TAL, ainsi que sur les techniques de d'analyse de dérivés et de composés.

tâches impliquées. Dans le contexte du projet ANR EDyLex, dans lequel ce travail a été réalisé, l'ambition était de permettre la mise en place d'outils d'indexation et de recherche automatiques destinés à faciliter le travail des journalistes, et notamment l'identification de thématiques émergentes, l'enrichissement de dépêches au moyen de métadonnées (informations biographiques sur les entités mentionnées dans les dépêches), l'amélioration de systèmes de transcription automatique de la parole, par exemple pour accélérer la publication de propos tenus par des personnages publics, ou encore l'accès à des dépêches traitant d'une thématique ou indiquant les prises de position d'un personnage public donné sur cette thématique. Outre le système évoqué ici, qui se concentre sur l'analyse des inconnus autres que les entités nommées, ces besoins nous ont conduit à travailler dans le cadre d'EDyLex, et entre autres sujets, sur la détection et plus encore sur la résolution d'entités nommées (Villemonte de La Clergerie *et al.*, 2009b ; Stern et Sagot, 2010a,b ; Béchet *et al.*, 2011 ; Stern *et al.*, 2012 ; Sagot et Stern, 2012 ; Stern et Sagot, 2012 ; Sagot *et al.*, 2012) [33] mais également, dans la suite de travaux réalisés dans le cadre du projet SCRIBO (Villemonte de La Clergerie *et al.*, 2009b) sur l'analyse linguistique et l'extraction automatique de citations (Sagot *et al.*, 2010 ; Danlos *et al.*, 2010) [34].

Nous nous sommes appuyé sur un corpus de dépêches de l'AFP collectées entre 2007 et 2013 et filtrées automatiquement afin d'en éliminer les tableaux de résultats sportifs, sommaires, agendas, signatures et autres éléments qui ne sont pas à proprement parler du contenu linguistique. Nous en avons sélectionné trois sous-parties afin de mener nos expériences : des dépêches entre le 24 juin et le 3 juillet 2009 (`AFP-annot`), l'intégralité des dépêches de l'année 2009 (`AFP-2009`) et 200 dépêches tirées au hasard entre le 1er et le 14 janvier 2013 (`AFP-eval`). Le tableau 3.3 donne les caractéristiques générales de ces corpus. Les corpus `AFP-annot` et `AFP-2009` sont utilisés à fins d'études. En particulier, `AFP-annot` avait été annoté manuellement en inconnus selon la classification de Blancafort San José *et al.* (2010) [35].

Outre les volumes de ces différents corpus, le tableau 3.3 renseigne sur le nombre d'inconnus qu'ils contiennent, tels que détectés par le module dédié de détection des inconnus au sein de notre chaîne de traitement décrite plus bas. Nous renvoyons à (Sagot *et al.*, 2013) pour un panorama qualitatif et quantitatif des différents types d'inconnus rencontrés dans ces données. Nous nous contenterons de donner ici quelques exemples : *corapporteur*, *anti-fraude* (dérivation préfixale), *bravitude*, *talibanisation* (dérivation suffixale), *politico-judiciaire*, *chiraco-villepinistes* (composition).

33. Ces travaux ont été effectués dans le cadre d'EDyLex et avant lui du projet SCRIBO (projet du pôle de competitivité System@tic) et conjointement dans celui de la thèse de Rosa Stern (Stern, 2015), thèse CIFRE en partenariat avec l'Agence France-Presse dont j'étais l'encadrant principal et dont Denis Teyssou était le responsable scientifique en entreprise.

34. Travaux effectués principalement avec Rosa Stern, alors doctorante CIFRE à l'Agence France-Presse et à Alpage, dont j'étais le co-encadrant principal.

35. Le travail d'annotation manuelle a été réalisé sous la responsabilité et avec les outils de l'entreprise Syllabs, dans le cadre du projet ANR EDyLex.

La figure 3.2 permet d'avoir un aperçu de l'évolution mois par mois et en fréquence relative de ces inconnus tels qu'identifiés au sein du corpus `AFP-2009`. Au fil de l'année, les accumuler permet d'indiquer pour chaque mois le nombre de nouveaux inconnus à traiter (*Nouveaux*), et leur intersection depuis le début de l'année (*Intersection*). Malgré les volumes de données que nous manipulons, il semble que le nombre de nouveaux inconnus apparaissant tous les mois diminue relativement peu. Cette apparition continuelle de nouvelles entrées, dans un corpus aussi contrôlé que des dépêches AFP, confirme la pertinence de la mise en place de mécanismes dynamiques pour les traiter.

Corpus	#Dépêches	#Tokens	#Inconnus	#Inconnus distincts
`AFP-annot`	2 535	1 060 378	6 208	2 782
`AFP-2009`	311 981	94 967 771	907 570	107 496
`AFP-eval`	200	73 353	729	489

Tableau 3.3 – Données quantitatives sur nos trois corpus de dépêches AFP, y compris le nombre d'inconnus qu'ils contiennent.

3.2.3 Architecture générale pour le traitement des inconnus

La figure 3.3 présente l'architecture générale que nous avons mise en place. Tous les modules ont été intégrés, pour certains de façon optionnelle, dans notre chaîne de traitement SxPipe (Sagot et Boullier, 2006, 2008 ; cf. chapitre 7).

Nous commençons par appliquer certains modules de la chaîne SxPipe qui effectuent la tokenisation du texte, la détection de motifs par automates (nombres, dates, sigles) et la reconnaissance d'entités nommées à l'aide de la base *Aleda* (Sagot et Stern, 2012) et de quelques motifs contextuels [36].

Un module dédié identifie alors les tokens inconnus et les étiquette comme tels, en prenant en compte les tokens absents du lexique de référence mais néanmoins analysables (tokens de type `donne-moi`). Le module suivant, décrit dans (Baranes, 2012), permet d'écarter les tokens correspondant à des emprunts à l'anglais non-adaptés [37, 38]. Parmi les inconnus restants, tous *candidats* à être des formes néologiques, nous avons cherché à repérer les formes qui auraient intérêt à être couvertes par le lexique, en quatre étapes décrites plus en détail dans (Sagot *et al.*, 2013) :

1. **Recherche dans des lexiques autre que le Le*fff***
   La notion d'inconnu étant définie ici par rapport à un lexique de référence, le

36. Parmi les options disponibles dans certains de ces modules, nous avons désactivé celles qui cherchent à corriger les fautes d'orthographe ou qui décomposent la reconnaissance des tokens (par dérivation ou par composition).

37. Des néologismes empruntés à des formes anglophones peuvent ne pas être repérés par ce module (*cardio-training*, *box-office*, etc.), mais ces erreurs représentent moins de 1% des tokens inconnus.

38. Ce travail a été effectué dans le cadre de la thèse de Marion Baranes, dont j'étais le principal encadrant.

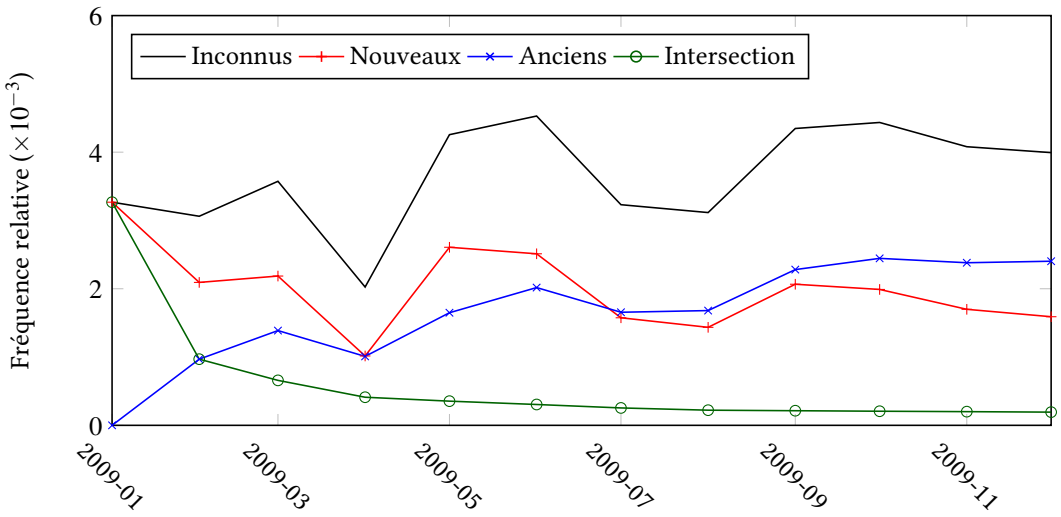


FIGURE 3.2 – Évolution temporelle des inconnus dans le corpus `AFP-2009`.

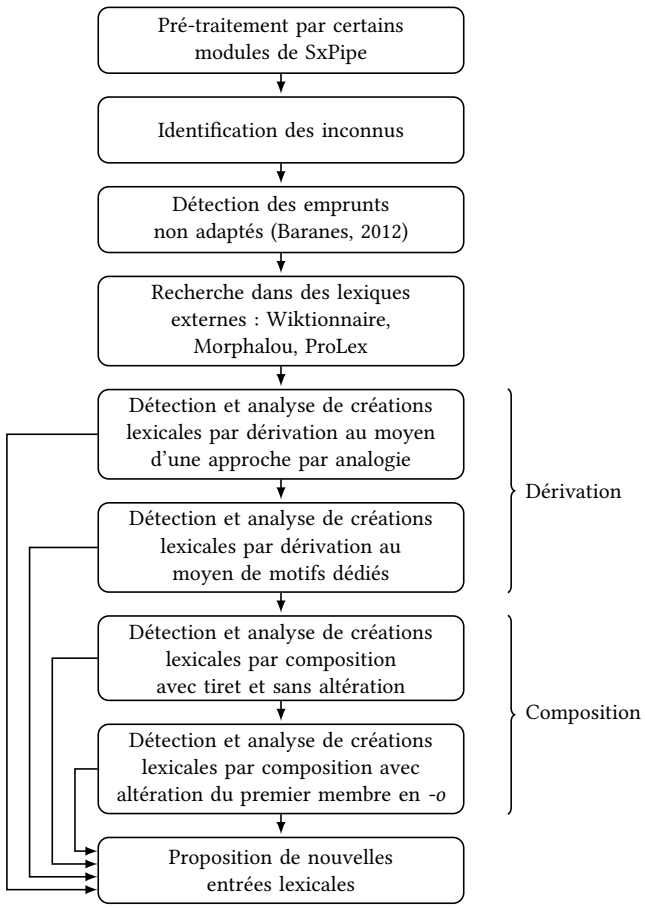


FIGURE 3.3 – Chaîne de traitement des inconnus

Le*fff*, la façon la plus simple de les traiter consiste à les rechercher dans d'autres ressources lexicales librement disponibles. Nous avons fait appel au Wiktionnaire, à Morphalou (Romary *et al.*, 2004), lexique morphologique extrait du TLFi, et à ProLexBase (Maurel, 2008), base de noms propres incluant de nombreux gentilés. Ces ressources ont été converties au format Alexina, de telle façon qu'elles soient représentées sous la forme de lexiques intensionnels faisant usage des mêmes classes flexionnelles que le Le*fff* (cf. section 3.1.2). Après élimination des entrées erronnées détectées par ce processus (cf. la remarque à ce sujet faite dans la section 3.1.3), Wiktionnaire, Morphalou et ProLexBase ont été ainsi transformés en des lexiques Alexina [39]. Les entrées lexicales ainsi obtenus, candidates à l'ajout dans le Le*fff*, couvrent au sein du corpus `AFP-2009` 18,6% des candidats distincts, parmi lesquels 71,5% sont trouvés dans le Wiktionnaire, 32,0% dans Morphalou et 14,9% dans ProLexBase (une entrée pouvant se trouver dans plusieurs lexiques en même temps).

2. **Détection et analyse par analogie des créations lexicales par dérivation**
   Le module décrit ici s'appuie sur les travaux présentés à la section 3.3.2 pour l'extraction de relations dérivationnelles entre entrées flexionnelles. Plus précisément, nous utilisons la même technique pour extraire des règles d'affixation communes à des paires d'entrées du Le*fff* qui en relient des formes fléchies à des formes de citation, mais dans le but, cette fois, d'obtenir des informations concernant un néologisme dérivationnel candidat. Ainsi, s'il est possible de relier un inconnu à une entrée du Le*fff* grâce à une règle de transformation de nature dérivationnelle, le lemme correspondant peut être obtenu par analogie avec les entrées lexicales du Le*fff*. L'application, en parallèle, de l'étiqueteur morphosyntaxique MElt (Denis et Sagot, 2012 ; cf. également chapitre 8) permet de ne conserver que les candidats dont la catégorie est la même que celle proposée par MElt pour l'inconnu. Nous avons ainsi pu analyser 11,9% des candidats extraits du corpus `AFP-2009`.

3. **Détection et analyse par règles des créations lexicales par dérivation**
   Notre étude préalable sur corpus nous a permis d'identifier certains phénomènes de création lexicale pour lesquels nous mettons au point des mécanismes d'analyse dédiés. En particulier, parmi les préfixes considérés, une proportion importante correspond à des dérivés affixaux de même catégorie et de même classe flexionnelle que leur base base. Ainsi, nous mettons en place un module qui, pour un inconnu donné, recherche un des préfixes standard et vérifie si la base candidate correspondante est un mot connu du Le*fff*. Dans le corpus `AFP-2009`, 16,6% des

39. Nous avons ainsi obtenu produisant environ respectivement 1 million, 400 000 et 125 000 entrées à partir du Wiktionnaire, de Morphalou et de ProLexBase, produisant au total 1 100 000 formes fléchies distinctes, parmi lesquelles 700 000 ne sont pas couvertes par le Le*fff*.

candidats distincts sont analysés grâce à ce module. Un mécanisme similaire a été mis en œuvre pour quelques suffixes mais ne traite qu'un très faible nombre d'inconnus (0,2%).

4. **Détection et analyse des créations lexicales par composition**
   Nous nous sommes restreints aux composés marqués typographiquement par un ou plusieurs tirets. Nous avons identifié, par ordre de préférence, (i) les mots composés dont le ou les espaces ont été remplacés par des tirets (exemple : *centre-ville*), les séquences de mots connus du Le*fff* et reliés par des tirets (exemple : *député-maire*), et, pour le candidat ne contenant qu'un seul tiret, ceux dont le dernier composant est un mot du Le*fff* et dont le premier se termine en *-o* et a un préfixe commun commun avec un mot du Le*fff* et couvrant au moins la moitié de sa longueur (exemple : *lumino-technique*). Pour tous les candidats ainsi détectés, nous construisons un lemme sur la base du dernier composant, dont nous avons observé qu'il imposait presque toujours sa catégorie et sa classe flexionnelle [40]. Parmi les candidats distincts traités par nos modules, celui-ci en analyse 57,7% dans le corpus `AFP-2009`.

Tous les modules que nous avons décrits, lorsqu'ils parviennent à analyser un inconnu, fournissent pour chaque élément son lemme, c'est-à-dire sa forme de citation, sa catégorie et sa classe flexionnelle. En conséquence, nous sommes en mesure de récupérer les analyses produites afin de proposer de nouvelles entrées à ajouter au lexique. L'ordre de ces modules importe : le premier qui parvient à analyser un inconnu interrompra le processus d'analyse. Il s'avère que la précision des modules d'analyse nous permet d'examiner les analyses dès la première occurrence.

3.2.4 Construction et évaluation d'entrées lexicales néologiques

Comme indiqué plus haut, l'objectif de ce travail est double : construire des entrées lexicales flexionnelles à ajouter au Le*fff* afin d'en augmenter la couverture sur les dépêches AFP de façon dynamique, mais également extraire des informations constructionnelles concernant ces nouvelles entrées, afin de permettre des traitements ultérieurs (sémantique lexicale y compris pour des applications en TAL, étude quantitative des mécanismes de création lexicale, etc.). Nous avons donc procédé à une évaluation en trois étapes, qui vise à répondre aux questions suivantes pour chaque occurrence d'inconnu : a-t-elle été correctement identifiée comme étant ou n'étant pas un néologisme ? si oui, l'entrée lexicale proposée est-elle correcte, y compris sa classe flexionnelle afin de pouvoir produire correctement ses formes fléchies ? si oui, les informations constructionnelles associées sont-elles correctes ?

40. En cas d'ambiguïté, nous utilisons l'étiquetage morphosyntaxique fourni par MElt pour choisir.

Pour cela, nous traitons les inconnus contenus dans le corpus `AFP-eval` tel que décrit à la section 3.2.3. Parmi les 489 inconnus distincts, 449 (soit 92%) ont été correctement classés, dont 357 qui ne sont pas des néologismes et 92 néologismes. Les 40 inconnus restant (8% du total) ont été mal classés, dont 34 néologismes : seulement 6 inconnus ont été analysés à tort comme des néologismes (et ont donc donné lieu à des entrées lexicales candidates erronées). Nous obtenons donc pour la tâche de détection des néologismes une précision de 94% et un rappel de 73%, et pour la tâche complémentaire de détection des inconnus non-néologiques une précision de 91% et un rappel de 98%.

Les 98 inconnus détectés comme néologismes, y compris les 6 classés par erreur, ont conduit à la création de 93 entrées lexicales candidates, c'est-à-dire d'entrées (forme de citation, catégorie, classe flexionnelle) qui permettent de construire automatiquement toutes les formes fléchies correspondantes. Nous avons évalué cette liste manuellement avec les résultats suivants : 73 sur 93, soit environ 80%, sont totalement correctes, 5 ont la bonne catégorie mais pas la bonne classe flexionnelle (ainsi *point-presse*, considéré comme féminin et prenant un *s* au pluriel), 13 n'ont pas la bonne catégorie (mais souvent les bonnes formes fléchies, car il s'agit fréquemment de confusions nom/adjectif, par exemple MULTI-FACETTE), 1 est douteuse et 1 est totalement erronée (le verbe MULTI-VOIR pour l'adjectif *multi-vues*).

Parmi les 73 entrées lexicales correctes, 52 ont été construites par l'un de nos modules d'analyse, et non au moyen de lexiques externes. Pour ces 52 entrées nous disposons donc d'informations constructionnelles, de nature à permettre le calcul d'informations supplémentaires telles que la valence ou la sémantique lexicale [41]. Nous avons étudié manuellement les informations constructionnelles obtenues au cours du processus d'analyse. Pour cette évaluation, celles-ci se sont toujours avérées correctes.

3.3 Développement de lexiques dérivationnels

On peut distinguer deux familles d'approches pour la prise en compte des mécanismes constructionnels [42] dans les travaux de TAL ou de linguistique quantitative. La première consiste à développer des lexiques constructionnels, c'est-à-dire des ressources comportant des relations constructionnelles entre entrées lexicales. Elle fait l'objet de cette section, dans laquelle nous nous restreindrons aux seules relations dérivationnelles. La seconde, que nous évoquerons à la section suivante, consiste à mettre en œuvre des systèmes d'analyse morphologique à la volée afin d'analyser des formes individuelles, par exemple des néologismes.

41. Ces informations pourraient également être construites pour les néologismes trouvés dans les lexiques externes.

42. La morphologie constructionnelle rassemble tous les mécanismes morphologiques de création lexicale.

Le développement de ressources lexicales dérivationnelles peut lui-même être envisagé de deux façons différentes, selon la manière dont l'on construit un modèle de la morphologie dérivationnelle de la langue considérée. Nous illustrerons successivement l'approche manuelle, dans laquelle les relations dérivationnelles ou les mécanismes dérivationnels qui leur sont sous-jacents sont formalisés manuellement *a priori* (cf. section 3.3.1), et un exemple d'approche automatique, où ces mécanismes sont induits à partir de données lexicales flexionnelles, avec différents degrés de supervision, notamment de par l'utilisation ou non d'informations complémentaires au développement parfois coûteux, comme des définitions lexicographiques (cf. section 3.3.2).

3.3.1 Acquisition automatique de liens dérivationnels à partir de règles de dérivation et de corpus bruts [43]

Le nombre de ressources lexicales contenant des relations dérivationnelles entre entrées lexicales reste limité. Si l'on se restreint aux quatre langues dont il sera question dans cette section, à savoir l'anglais, l'allemand, l'espagnol et le français, seul ce dernier semble avoir bénéficié d'efforts importants dans cette direction — et ce malgré notamment le lexique CELEX (Burnage, 1990), cité plus haut, qui couvre l'anglais, l'allemand et le néerlandais. Ainsi, *VerbAction* (Tanguy et Hathout, 2002) associe des verbes avec leurs noms d'action dérivationnellement reliés (⟨*accuser*, *accusation*⟩), et *VerbAgent* (Tribout *et al.*, 2012) avec leurs noms d'agent dérivationnellement reliés (⟨*accuser*, *accusateur*⟩). D'autres travaux partent des noms et les associent aux verbes qui leur sont dérivationnellement reliés, et notamment le lexique Nomage (Balvet *et al.*, 2011) pour le français ou Nomlex (Macleod *et al.*, 1998) pour l'anglais. La base Mordan (Koehl, 2013) contient plusieurs milliers de paires formées de noms désadjectivaux et de leur base adjectivale. On peut également citer Polymots (Gala *et al.*, 2010), une ressource lexicale qui regroupe les mots en familles morphologiques et fournit quelques informations complémentaires. Ces ressources ont été le plus souvent développées de façon manuelle ou à partir de d'informations lexicales développées manuellement (descriptions manuelles des mécanismes en cause, définitions lexicographiques, etc.). En particulier, l'ajout manuel de règles de dérivation à une description morphologique permet des descriptions fines, qui prennent en compte les procédés peu fréquents et leurs différentes variantes, mais peut s'avérer laborieuse et doit être répétée pour chaque nouvelle langue.

C'est cette approche que nous avons mise en œuvre lors du développement du lexique de noms déverbaux DeNALex, dont le développement est présenté ci-dessous (cf. également Strnadová et Sagot, 2011). En effet, en dépit des ressources mentionnées

43. Ce travail a été en collaboration avec Jana Strnadová, alors doctorante au Laboratoire de Linguistique Formelle (LLF) sous la codirection de Bernard Fradin (LLF, CNRS) et de Pavel Štichauer (Université Charles à Prague, République tchèque). Il a fait l'objet d'une publication (Strnadová et Sagot, 2011), le travail de validation ayant été poursuivi par après.

ci-dessus, il semble qu'il n'existait pas pour le français, avant le travail décrit dans cette section, de lexique d'adjectifs dénominaux (la contraposée de la base MORDAN mentionnée ci-dessus, qui traite des noms désadjectivaux). Nous avons donc cherché à construire une telle ressource, en nous concentrant sur les adjectifs dérivés à partir d'une base nominale par affixation régulière, tout en prenant en compte les nombreuses variantes possibles selon les affixes et les bases. Nous renvoyons à (Strnadová et Sagot, 2011) pour une description des adjectifs dénominaux en français ainsi que pour une description de l'approche suivie, que l'on peut résumer en trois étapes :

1. Ajout dans la grammaire morphologique du Le*fff* des opérations dérivationnelles possibles, pour certaines contraintes par des propriétés formelles du radical de la base nominale ; naturellement, on ne s'attend pas, loin de là, à ce que tous les adjectifs dérivés rendus ainsi compatibles avec la grammaire soient effectivement en usage dans la langue ;
2. Induction automatique de liens de dérivation formels entre entrées lexicales connues de lexiques flexionnels de deux lexiques de référence, à savoir le Le*fff* ainsi que Morphalou après transformation en un lexique Alexina faisant usage des mêmes classes flexionnelles que le Le*fff* (cf. section 3.1.2) ; un lien entre une entrée nominale et une entrée adjectivale est créé dès lors que ce lien est formellement permis par la grammaire morphologique étendue au point précédent [44] ; nous obtenons ainsi 3 293 liens dérivationnels distincts (cf. table 3.4) ;
3. Induction automatique de liens de dérivation formelle entre noms connus et dérivés adjectivaux inconnus mais attestés dans un corpus volumineux de 5 milliards de tokens, créant ainsi des candidats entrées lexicales nouvelles ; l'idée générale consiste à construire tous les adjectifs dérivés potentiels à partir de l'ensemble des noms du corpus, puis à ordonner ceux d'entre eux qui ne sont pas présents dans le lexique flexionnel de départ en fonction de leur attestation ou non dans le corpus ; enfin, les couples (lemme nominal de base, lemme adjectival dérivé inconnu des lexiques flexionnels) sont filtrés ; sont conservés 7 449 nouveaux lemmes adjectivaux inconnus, reliés à des bases nominales connues par 8 736 liens dérivationnels.

Nous avons estimé la précision et la couverture des résultats obtenus par différents moyens décrits dans (Strnadová et Sagot, 2011). La précision est d'environ 85 % pour les adjectifs dénominaux connus du lexique et de 51 % pour ceux qui sont inconnus — dans

44. L'union des entrées nominales du Le*fff* et de Morphalou comporte 65 651 lemmes (forme citationnelle et classe flexionnelle), qui produisent par ces règles un total de 886 526 couples (nom base, adjectif dérivé) candidats. Parmi ces couples, 3 293 ont un adjectif dérivé qui est lui aussi du Le*fff* ou de Morphalou. Ils concernent 2 687 adjectifs distincts, un même adjectif pouvant être obtenu par dérivation à partir de plusieurs bases (correctes ou non). Ces couples sont donc des relations dérivationnelles candidates entre noms et adjectifs déjà connus du lexique.

Type de couple (nom base, adjectif dérivé)	#couples	#adj. distincts
Candidats produits	**886 526**	**844 519**
1. retenus car l'adjectif est connu du lexique (Le*fff*+Morphalou)	**3 293**	**2 687**
2. à adjectif inconnu mais retenus après confrontation au corpus		
avant filtrage	12 140	10 064
après filtrage des candidats proches de mots connus	11 463	8 317
après filtrage des couples à nom ou adjectif trop court	**8 736**	**7 449**
Total des candidats retenus	***12 029***	***9 692***

Tableau 3.4 – Résultats quantitatifs de l'extraction de couples (base nominale, adjectif dérivé). On note qu'en plus de construire des relations de dérivation morphologique au sein du lexique, notre approche a permis l'identification de 7 449 candidats néologismes, termes techniques et autres types d'adjectifs absents et du Le*fff* et de Morphalou.

tous les cas, leur base nominale est connue, par construction. Le rappel, quant à lui, est proche de 100 % si l'on se restreint aux adjectifs dont la base nominale est connue.

La ressource obtenue alors été validée et corrigée manuellement. Le résultat, le lexique DeNALex (*De-Nominal Adjective Lexicon*) est librement disponible sous licence LGPL-LR comme complément au lexique Le*fff*. Une perspective naturelle, grâce notamment à des moyens complémentaires comme des lexiques bilingues français–latin, une modélisation partielle de l'évolution diachronique du lexique du français ou encore des modèles sémantiques distributionnels, serait de compléter DeNALex par des couples (adjectif dérivé, base nominale) relevant de la dérivation à base supplétive, dont nous avons pu estimer au cours de ce travail qu'elle concernait une petite moitié des adjectifs dénominaux. C'est en effet l'avantage principal de ce type d'approches que de pouvoir procéder à des modélisations fines des mécanismes en jeu, afin de réussir à extraire des relations dérivationnelles difficilement accessibles par des moyens moins supervisés.

3.3.2 Acquisition automatique endogène de liens dérivationnels dans un lexique flexionnel [45]

Au vu du travail manuel linguistique important que nécessitent les approches telles que celle présentée à la section précdente, plusieurs auteurs ont proposé de développer des ressources dérivationnelles de façon moins supervisée, parfois non supervisée. Une telle approche ne garantit ni la couverture des cas rares ou des variantes spécifiques ni la qualité des règles dérivationnelles extraites. Du reste, les règles dérivationnelles que l'on peut obtenir automatiquement ne sont pas toujours de règles de dérivation *stricto sensu*, mais plutôt de *motifs de transfert* permettant de passer d'un lemme à un autre au sein

45. Ce travail a été en collaboration avec Marion Baranes dans le cadre de sa thèse (Baranes, 2015), thèse dont j'étais encadrant principal et co-directeur. Il a fait l'objet d'une publication (Baranes et Sagot, 2014a).

d'une même *famille morphologique*[46]. Une approche non supervisée permet cependant de couvrir très rapidement l'ensemble des liens dérivationnels les plus fréquents dans le lexique, sans travail supplémentaire pour traiter une nouvelle langue. Can et Manandhar (2009) décrivent ainsi une approche non supervisée qui repose, entre autres, sur les parties du discours afin de produire une analyse morphologique. Bernhard (2010) décrit deux systèmes non supervisés distincts, *MorphoClust* et *MorphoNet*. Le premier repose sur une technique de *clustering* hiérarchique pour regrouper les formes d'une même famille au sein d'un réseau lexical, alors que le second repose sur un algorithme de graphe. Walther et Nicolas (2011) utilisent un segmenteur morphologique non supervisé pour identifier des morphes dérivationnels et en déduire des relations dérivationnelles au sein d'un lexique flexionnel. Une famille d'approches utilisées régulièrement dans ce cadre repose sur la notion d'analogie (Lepage, 1998, 2000 ; Stroppa et Yvon, 2005, 2006 ; Lavallée et Langlais, 2009, 2011). Hathout (2010) combine plusieurs de ces approches pour développer Morphonette, un système restreint au français qui repose à la fois sur l'analogie et sur des notions de similarité morphologique et de similarité sémantique (cette dernière reposant sur des définitions lexicographiques). Une telle combinaison, si elle permet d'améliorer la qualité et la richesse des résultats, se fait au détriment de l'intérêt principal des approches non supervisées : leur applicabilité à un large ensemble de langues sans développement spécifique important. C'est notamment le cas dès lors que l'on fait usage de définitions lexicographiques.

Ainsi, nous avons également exploré les approches automatiques permettant l'extraction de relations dérivationnelles au sein de lexiques flexionnels (Baranes et Sagot, 2014a). Plus spécifiquement, nous avons développé un système destiné à extraire de façon non supervisée, au moyen d'une approche par analogie, des règles de transformation pondérées qui relient des paires d'entrées lexicales flexionnelles dérivationnellement reliées[47] (cf. Gaussier, 1999), et utilise ces règles pour construire des liens dérivationnels à partir d'un lexique purement flexionnel. Nos règles de transformation peuvent également être utilisées pour acquérir des informations morphologiques (flexionnelles et dérivationnelles) pour des formes (mots) trouvées en corpus et inconnues du lexique flexionnel. Notre système est indépendant de la langue, bien que restreint à ce stade à la morphologie dérivationnelle concaténative. Nous l'avons évalué sur quatre langues, l'anglais, le français, l'allemand et l'espagnol.

Notre système repose sur la notion d'analogie : pour identifier au sein du lexique flexionnel de départ des lexèmes appartenant à une même famille morphologique, nous cherchons des règles affixales qu'ils partagent. Par exemple, nous pourrons déduire que

46. Nous définissons une famille morphologique comme un ensemble de lemmes reliés les uns aux autres par des relations morphologiques dérivationnelles.

47. Nous qualifions de *dérivationnellement reliées* deux entrées lexicales flexionnelles (lemmes) qui correspondent chacune à un lexème appartenant à la même famille morphologique.

l'adjectif anglais *liable* appartient à la même famille morphologique que *liability* si l'on relève que : (1) *liability* est connu (il fait partie du lexique morphologique de départ), et (2) il existe une règle qui nous permet de substituer le suffixe *–ability* à *–able*. Nous commençons donc par une étape d'extraction de telles règles, que nous appellerons *règles de transformation* [48], de façon non supervisée. Nous commençons par extraire des règles reliant des formes fléchies à des formes de citations (cf. Baranes et Sagot, 2014a pour plus de détails). Des exemples de règles retenues sont donnés à la table 3.5. Une fois extrait ce jeu de règles de transformation, nous l'utilisons pour relier les entrées lexicales de notre lexique flexionnel, *via* les couples (formes fléchies, formes de citation) que ces règles autorisent à construire [49].

Langue	Catégorie	Préfixe	Suffixe	Occ	Exemple
anglais	A_ → R_inf	_ → *un-*	_ → *-ly*	1123	*fortunate* → *unfortunately*
allemand	adj_plain.pl.nom.primary.long → n_sg.gen.short	—	*-ische* → *-ie*	136	*morphologische* → *morphologie*
espagnol	v_MN0000 → n_CMS000	*a-* →_	*-ar* → *-o*	342	*abalear* → *baleo*
français	adj_Kfp → v_W	_	*-ées* → *-er*	6483	*données* → *donner*

Tableau 3.5 – Exemples de règles de transformation extraites des différents lexiques flexionnels. La colonne « Catégorie » illustre les informations manipulées, qui incluent à la fois la partie du discours et l'étiquette morphologique.

Appliqué aux lexiques flexionnels Alexina (ou Alexina$_{\text{PNRSLI}}$) de l'anglais, de l'allemand, de l'espagnol et du français, en nous restreignant aux entrées adjectivales, adverbiales, nominales et verbales, nous avons construit des lexiques dérivationnels dont les caractéristiques quantitatives sont indiquées à la table 3.6 [50].

Nous avons évalué notre système en deux étapes. Nous avons tout d'abord mesuré la précision de nos couples d'entrées, en en évaluant manuellement un sous-ensemble aléatoirement choisi de 100 couples par langue. Les taux d'erreur pour l'anglais, le français et l'allemand sont très bas (de 2% à 6%), bien qu'ils soient à prendre avec prudence étant

48. Ces règles de transformation, contrairement à celles du paragraphe précédent, développées manuellement, ne sont pas nécessairement des règles de dérivation. Elles modélisent des transformations qui relient deux lexèmes appartenant à la même famille morphologique, qu'ils soient reliés dérivationnellement ou non, et indépendamment de l'éventuelle directionnalité d'un tel lien de dérivation.

49. On notera qu'un lexique purement flexionnel ne fait pas de distinction entre les différents sens d'un même flexème (l'emploi du mot « lexème » dans ce qui précède est donc abusif). Il se peut que l'on relie des entrées lexicales par un lien qui, en réalité, ne s'applique qu'à des sens (ou lexèmes) particuliers des flexèmes concernés, la notion de famille morphologique incluant en effet une notion d'affinité sémantique.

50. Pour plus de détails sur les règles extraites, y compris pour une discussion sur le nombre de couples extraits et leur répartition en fonction des catégories de départ et d'arrivée, on pourra se reporter à (Baranes et Sagot, 2014a).

Lexique	#Règles de transformation	#Couples d'entrées	#Paires extraites / #Paires possibles
anglais	11 748	597 148	0,015‰
allemand	6 812	10 639	0,017‰
espagnol	6 000	69 694	0,005‰
français	8 834	84 927	0,003‰

Tableau 3.6 – Nombre de règles de transformations et de couples d'entrées lexicales extraits

donnée la taille des échantillons, mais le taux d'erreur pour l'espagnol est plus élevé (13%) [51].

Afin d'estimer le rappel, nous avons procédé à une seconde évaluation, restreinte au français. Nous avons comparé nos couples d'entrées lexicales à la ressource Morphonette (Hathout, 2010), dont la construction repose elle aussi sur l'analogie, mais qui a bénéficié d'informations lexicographiques développées manuellement (cf. plus bas). Une fois éliminées de Morphonette toutes les relations flexionnelles ainsi que les relations impliquant des composés néoclassiques et non couvertes par notre approche (par exemple la forme *psychopathe*), les couples de Morphonette retenus sont au nombre de 76 754 (sur les 96 081 couples de la ressource dans son ensemble). Les deux jeux de données ont alors 22 591 couples en commun, 62 336 couples n'étant créés que par notre système et 54 163 couples ne se trouvant que dans Morphonette. Notre système et Morphonette ont donc environ un tiers de couples en commun, ce qui est relativement peu : le rappel des deux approches reste améliorable. Si l'on considère l'union des deux jeux de données comme une référence pour le calcul du rappel, ce qui est naturellement optimiste, on obtient des valeurs de rappel de 61% pour notre système et de 51% pour Morphonette.

Pour comparer ces trois ensembles de couples, nous avons évalué manuellement pour chacun d'entre eux 200 paires choisies aléatoirement. Les couples trouvés à la fois dans Morphonette et par notre système ont un taux de précision de 97,5%, ceux trouvés uniquement par notre méthode ont une précision de 94%, et ceux présents uniquement dans Morphonette sont précis à 96,5%. La précision de notre approche donc est quasiment aussi élevée que celle de Morphonette, ce qui est satisfaisant pour au moins trois raisons. Premièrement, nous construisons plus de couples que Morphonette. Deuxièmement, il est important de rappeler que la construction de Morphonette repose sur des informations lexicographiques riches construites manuellement, puisqu'il a été fait usage des définitions contenues dans la version électronique du *Trésor de la Langue Française*. Malgré le caractère totalement non supervisé de notre approche, la précision

51. Ceci résulte en grande partie de la présence d'une règle bruitée qui crée des liens dérivationnels entre lexèmes qui se différencient par les suffixes *-ear* et *-ar*. Cette règle crée des couples erronés tels que ⟨*zapar 'saper'*, *zapear 'chasser, éloigner (un chat)'*⟩ ou ⟨*copear 'boire (familier)'*, *copar 'accaparer (le marché)'*⟩. Cette règle est à l'origine de 9 des 13 couples étiquetés err.

que nous obtenons est quasiment aussi élevée. Enfin, notre système est indépendant de la langue, dans les limites indiquées plus haut, et nous en avons effectivement montré l'efficacité sur quatre langues différentes, alors que développer des équivalents de Morphonette pour d'autres langues nécessiterait d'utiliser de dictionnaires électroniques pour chacune de ces langues.

Toutefois, avec une approche comme celle présentée à cette section, il est bien plus difficile de faire émerger des relations dérivationnelles qui ne résultent pas d'opérations strictement ou quasiment concaténatives. Nous avons mentionné les dérivés néo-classiques, dont certains sont présents dans Morphonette, et qui ne sont pas couverts par notre approche. Pour de tels cas, des informations supplémentaires doivent être mises en œuvre, soit par le développement de lexiques de bases supplétives (associant par exemple *canin* à *chien*), soit par l'exploitation d'informations exta-morphologiques, comme c'est le cas des définitions lexicographiques dans Morphonette.

Chapitre 4

Morphologie quantitative

Sommaire

Ces dernières années, le développement de méthodes de mesure de la complexité linguistique est devenu un domaine de recherche actif, avec plusieurs objectifs, dont notamment la comparaison des langues entre elles, et notamment celle des créoles aux autres langues (McWhorter, 2001 ; Bonami *et al.*, 2011) et l'étude de la répartition de la complexité linguistique entre différentes sources ou niveaux d'analyse (Juola, 1998 ; Ehret, 2014). Par ordre décroissant de généralité, des travaux ont été réalisés dans ce domaine sur les langues prises de façon globale (McWhorter, 2001 ; Juola, 2008 ; Moscoso del Prado Martín *et al.*, 2004 ; Nichols, 2009), en restreignant le travail à un niveau d'analyse particulier tel que la morphologie (Bane, 2008) ou encore en mesurant la complexité de descriptions morphologiques particulières, notamment dans le contexte de l'acquisition

automatique non supervisée ou faiblement supervisée de la morphologie (Goldsmith, 2001 ; Xanthos, 2008).

L'interêt pour les études sur la complexité linguistique est alimenté par les différentes perpectives qu'ouvre cette notion, lesquelles motivent en retour des définitions ou des propriétés qui lui sont associées. Parmi ces perspectives, nous nous concentrerons sur la suivante : toutes choses égales par ailleurs, une modélisation d'un jeu de données est d'autant meilleure — quoi que cela veuille dire — que sa complexité est plus faible. L'objectif principal est ainsi de produire de meilleurs modèles de systèmes linguistiques. On le voit, bien que l'idée générale semble intuitive, sa mise en œuvre requiert une compréhension correcte de ce que l'on entend par « meilleure modélisation », condition préalable à une formalisation adaptée de la notion de complexité. On peut alors envisager de caractériser, pour une langue donnée, les propriétés des modélisations identifiées comme étant les meilleures, au moins parmi un espace de modélisations possibles que l'on se donne au départ.

Dans cette section, comme au chapitre 4, nous nous concentrerons sur la notion de *complexité morphologique*, et plus spécifiquement de *complexité flexionnelle* ; nous utiliserons par abus de langage l'un et l'autre termes de façon équivalente. Plus précisément, et conformément à ce que nous avons indiqué ci-dessus, nous nous pencherons sur différentes façons de définir et de mesurer ce que l'on peut vouloir dénoter par le terme de complexité morphologique : qu'est-ce que l'on veut dire si l'on affirme intuitivement que le système flexionnel d'une langue est plus complexe que celui d'une autre, et comment quantifier cette intuition ? Sans surprise, de nombreuses définitions de cette notion ont été proposées dans la littérature. C'est au point que reste ouverte la question de savoir ce que la notion de complexité morphologique pourrait ou devrait recouvrir, et *in fine* l'acception qu'il convient de donner au terme *complexité* dans ce contexte. Cette disparité est due en partie à ce que la notion de complexité morphologique, très vague dans l'absolu, peut être identifiée à différentes dimensions des propriétés des systèmes flexionnels ou de leur description, et par conséquent différentes façons de considérer ce qu'est une modélisation de ces systèmes et *a fortiori* ce qu'est une modélisation meilleure qu'une autre [1].

Nous nous concentrerons sur deux définitions possibles — et deux mesures associées — de la complexité morphologique, un panorama plus complet des approches existantes, et de leurs limites, étant proposé à la section A.4. Les deux définitions que nous avons retenues reposent toutes deux, quoique de façon différente, sur la théorie de l'information. La première part de descriptions explicites d'un système morphologique,

1. Une autre problématique, que nous laisserons pour partie de côté, est de déterminer quelles sont les limites du système flexionnel. Ainsi, nous reviendrons plus bas sur la place de la (morpho)phonologie dans ces travaux sur la complexité morphologique, mais nous laisserons de côté la question de savoir si les formes périphrastiques pourraient être prises en compte dans les calculs. Nous admettons une réponse négative à cette dernière question.

et fait l'hypothèse que la complexité d'un tel système peut être appréhendée par la mesure de la complexité d'une description idéale dudit système, cette dernière pouvant être estimée par la mesure de son contenu informationnel (sa *longueur de description*). Autrement dit, un système morphologique serait d'autant plus complexe qu'il est difficile de le décrire de façon compacte. On voit la limite, mais également l'intérêt, d'une telle définition : il est fort difficile d'être assuré de disposer d'une description idéale, et l'on sera de toute façon contraint par le formalisme utilisé pour la construire. Mais on voit aussi, comme nous l'avons proposé, que l'on peut ainsi disposer de la sorte d'un moyen objectif de comparaison entre descriptions concurrentes d'un même système (Section 4.1). Une seconde définition de la complexité morphologique s'appuie sur la prédictabilité des formes entre elles : elle fait l'hypothèse qu'un système morphologique est d'autant plus complexe qu'il est difficile de prédire l'ensemble du paradigme à partir d'une (dans certains travaux, de plusieurs) formes connues. Cette intuition se formalise également au moyen de la théorie de l'information, cette fois-ci *via* la notion d'entropie. La première mise en œuvre de cette approche par Ackerman *et al.* (2009) n'allait pas sans problème, et nous avons proposé une nouvelle piste (Section 4.2). Nous avons également travaillé à la combinaison de ces deux approches, dans le but de réduire la part d'arbitraire dans la compréhension des systèmes morphologiques : nous avons montré que la combinaison de la notion de longueur de description et de celle d'entropie entre cases du paradigme permet de faire émerger de façon endogène, sans description *a priori*, une structuration en classes flexionnelles (Section 4.3).

4.1 Complexité morphologique et descriptions concurrentes [2]

Comme indiqué ci-dessus, la complexité d'un système morphologique peut être appréhendée, avec un point de vue constructif, par la mesure de la longueur de description d'une modélisation de ce système morphologique. Or nous avons montré au chapitre 2 (et un peu plus en détail à l'annexe B) comment notre formalisme morphologique Alexina$_{\text{PARSLI}}$, qui implémente le modèle PARSLI de la morphologie flexionnelle de Walther (2016), permet de représenter les phénomènes flexionnels non canoniques. Différentes analyses concurrentes des mêmes données peuvent alors être implémentées en Alexina$_{\text{PARSLI}}$ et ainsi comparées quantitativement au moyen d'une mesure adaptée de leur longueur de description. Nous avons donc développé une telle mesure, décrite

2. Les travaux présentés par cette section ont été réalisés en collaboration avec Géraldine Walther (Université de Zürich), alors doctorante au Laboratoire de Linguistique Formelle, Université Paris Diderot, sous la direction d'Anne Abeillé (LLF, Université Paris-Diderot) et d'Olivier Bonami (LLF, Université Paris-Sorbonne). Ceux concernant le khaling (section 4.1.2) ont été obtenus en collaboration avec Géraldine Walther et Guillaume Jacques (CRLAO, CNRS), dans le cadre de l'opération LR 4.11 de l'axe 6 « Ressources linguistiques » du LabEx EFL, axe dont j'avais la responsabilité. Les publications associées sont citées au fil du texte.

en détail notamment dans (Sagot et Walther, 2011 ; Walther et Sagot, 2011a), avec deux objectifs fortement interconnectés : évaluer l'impact de l'utilisation de notions spécifiquement développées dans ᚠᚨᚱᛊᛚᚾ pour encoder les phénomènes non-canoniques, et évaluer l'impact de certains choix descriptifs. Nous décrirons ici deux études, l'une sur la flexion verbale du français (Sagot et Walther, 2011 ; Walther et Sagot, 2011a) [3] et l'autre sur celle du khaling (kiranti, sino-tibétain, Népal ; Walther *et al.*, 2013, 2014b) [4].

4.1.1 Mesurer la complexité de différentes descriptions de la flexion verbale du français : l'équilibre entre lexique et grammaire

La flexion verbale du français est intéressante à plusieurs égards, dont certains ont été mentionnés dans les chapitres précédents. Tout d'abord, il s'agit d'un système riche qui génère des formes pour une cinquantaine de structures de traits morphosyntaxiques (8 temps, 3 personnes et 2 nombres pour les formes fléchies hors formes de l'impératif, plus les formes de l'impératif, l'infinitif et les formes participales). Par ailleurs, ce système est traditionnellement décrit comme comportant une classe flexionnelle régulière et productive, celle des verbes du premier groupe (verbes en *-er*), une classe flexionnelle irrégulière (verbes du troisième groupe), et la classe flexionnelle des verbes du deuxième groupe (verbes comme *finir*), parfois considérée comme régulière et parfois comme irrégulière. Les analyses diffèrent sur la productivité réelle de cette dernière classe (Boyé, 2000 ; Kilani-Schoch et Dressler, 2005 ; Bonami *et al.*, 2008), ce qui constitue un premier point possible de divergence entre analyses différentes de la flexion verbale du français.

Par ailleurs, la flexion verbale du français est le lieu de différents phénomènes non-canoniques, tels que la surabondance régulière des verbes en *-ayer* sur une partie des paradigmes (*je balaie/balaye* ; cf. 2.2.1.4), l'existence de formes en *-i-* et de forme en *-iss-* dans les paradigmes des verbes du deuxième groupe, que l'on peut analyser comme deux radicaux ou en intégrant *-ss-* dans certains des suffixes, ou encore la multiplicité variable des radicaux pour les verbes du troisième groupe. Face à de tels phénomènes, plusieurs approches sont possibles, et qui se distinguent notamment par la distribution de l'information entre composantes lexicale et grammaticale de la description, ainsi que par l'utilisation ou non de certains mécanismes descriptifs tels que les règles morphographémiques ou les outils formels disponibles dans Alexinaᚠᚨᚱᛊᛚᚾ pour modéliser explicitement les phénomènes non-canoniques.

Nous avons retenu quatre approches, que nous avons implémentées ou réimplémentées en Alexinaᚠᚨᚱᛊᛚᚾ afin de pouvoir les comparer. À une extrémité du spectre, nous avons

3. Expérience effectuée sur une ancienne version d'Alexinaᚠᚨᚱᛊᛚᚾ et donc de la mesure de complexité associée.

4. Walther (2013b) a également utilisé cette mesure et nos outils pour des expériences sur le latin (italique, indo-européen) et sur le maltais (sémitique, afro-asiatique, Malte ; cf. aussi Camilleri et Walther, 2012). Une mise en perspective de l'ensemble de ces travaux a été proposée par Walther (2013a).

réimplémenté le modèle de Bonami et Boyé (2003) ; Bonami *et al.* (2008), lequel s'appuie sur une distribution des cases en 12 radicaux distincts, d'une modélisation de la façon dont on peut prédire chacun de ces radicaux à partir de tel ou tel autre, et d'un schème flexionnel unique (analyse BoBo) [5]. À l'autre extrémité du spectre, nous avons développé une analyse au moyen de l'algorithme simple décrit à la section 3.1.2 (analyse Flat), qui s'appuie pour chaque verbe sur un radical unique mais multiplie le nombre de schèmes flexionnels (139) et n'utilise ni règles morphographémiques ni aucun des outils formels disponibles dans Alexina$_{\text{PARSLI}}$ pour modéliser la non-canonicité. Entre les deux, nous avons retenu la description originelle du Le*fff*, qui utilise un grand nombre de règles morphographémiques ainsi que certains mécanismes de factorisation de la description, et nous avons développé une nouvelle description qui, outre un nombre raisonnable de règles morphographémiques, fait pleinement usage de la notion de *zone flexionnelle* définie dans PARSLI (20 schèmes flexionnels ; analyse new). En PARSLI, et donc en Alexina$_{\text{PARSLI}}$, une zone flexionnelle est un ensemble de cases du paradigme associé à un ensemble de règles de réalisation qui concernent l'un des niveaux d'analyse définis dans la description, et dont l'un doit être de type radical. Au niveau radical, on parle de *zone radicale*, un concept qui est équivalent à la notion d'*espace thématique* de Bonami et Boyé, 2003. Aux niveaux définissant les exposants, on parle de *zone d'exponence*. Illustrons ces notions sur l'utilisation qui en est faite dans la description New pour représenter la surabondance des verbes en *-ayer*. Dans cette description, nous avons défini deux zones radicales, l'une, en *-ay-*, qui couvre l'ensemble des structures de traits morphosyntaxiques exprimables par un verbe standard du français, et l'autre, en *-ai-*, qui ne concerne que le sous-ensemble pour lequel les verbes en *-ayer* sont surabondants. On peut alors produire toutes les formes du paradigme d'un verbe en *-ayer* au moyen d'un schème flexionnel rassemblant deux *sous-schèmes flexionnels* (combinaisons de zones flexionnelles mises en œuvre pour construire les formes du paradigme) : chaque sous-schème repose sur l'une de ces deux zones radicales et la combinent avec la zone d'exponence unique commune à tous les verbes du premier groupe.

Le développement ces quatre descriptions, illustrés par quelques entrées dans le tableau 4.1, s'est appuyé sur les données du Le*fff*, soit, dans la version utilisée, 7 820 verbes parmi lesquels (6 966 verbes du premier groupe, 315 du deuxième groupe et 539 du troisième groupe) produisant 300 693 formes fléchies distinctes. La figure 4.1 représente les résultats obtenus, en indiquant la distribution de la quantité d'information entre composantes lexicale et grammaticale des descriptions. On constate que l'utilisation des notions spécifiques à Alexina$_{\text{PARSLI}}$, et en l'espèce celle de zone flexionnelle, conduit à une compacité meilleure qu'avec les autres analyses. C'est le cas non seulement par rapport

5. L'analyse originelle a dû être étendue de différentes façons, notamment pour adapter à une analyse graphémique l'analyse originale phonémique (ce qui a conduit à l'adjonction de 61 règles morphographémiques) ainsi que pour gérer correctement les phénomènes de surabondance et de défectivité.

Forme de citation	Schème flexionnel associé			
	Flat	Orig	New	BoBo
aimer	v1	v-er$_{std}$	v-er	v_1
acheter	v18	v-er$_{std}$	v-er	v_1
jeter	v8	v-er$_{dbl}$	v-eter	$v_1/_{,jett,,,,,jett\ə}$
balayer	v12	v-ayer	v-ayer	v-ayer$_1$
finir	v2	v-ir2	v-ir2	v_{23r}
requérir	v42	v-ir3	v-ir3	$v_{23r}/_{requér,requier,,,,,requer,requi,requis}$
cueillir	v51	v-assaillir	$v_{23r}/_{cueill,,,,,,cueillə}$	$v_{23r}/_{cueill,,,,,,cueillə}$
prendre	v24	v-prendre	v-prendre	v_{3re}
mettre	v17	v-mettre	v-mettre	$v_{3re}/_{,met,,,,,,mi,mis}$

Tableau 4.1 – Entrées lexicales pour quelques lemmes dans chacune de nos quatre représentations concurrentes de la flexion verbale du français

aux descriptions traditionnelles mais aussi par rapport à la description récente et originale de Bonami *et al.* (2008) [6].

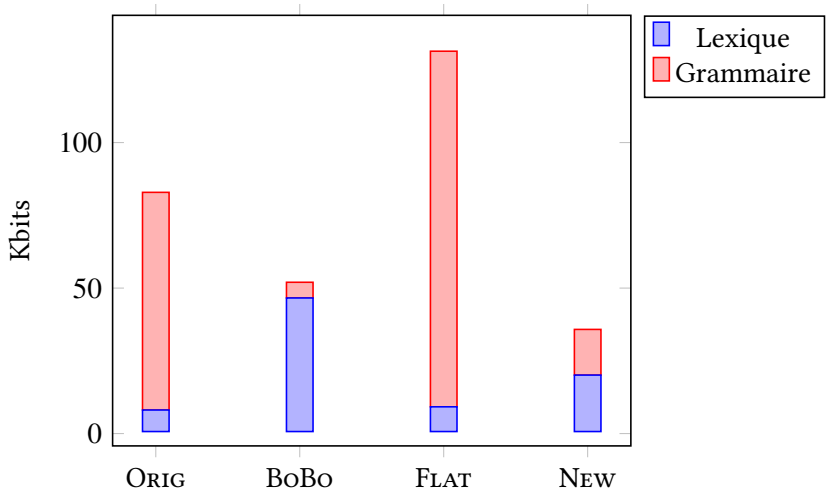


Figure 4.1 – Longueurs de description de différentes descriptions de la morphologie verbale du français.

4.1.2 Mesurer la complexité de différentes descriptions de la flexion verbale du khaling : la pertinence des traits morphomiques

Nous avons également mené des expériences consistant à comparer au moyen de cette mesure de complexité deux descriptions concurrentes de la morphologie verbale du khaling (kiranti, sino-tibétain, Népal ; Walther *et al.*, 2013, 2014b). Le khaling et une

6. On peut noter que nous n'aurions pas obtenu ce résultat si nous n'avions pas pris en compte la longueur de description du lexique, mais seulement celle de la grammaire morphologique. Cela dit, puisque la répartition de l'information morphologique entre lexique et règles varie d'une description à une autre, cela n'aurait pas grand sens d'évaluer la longueur de description de la grammaire morphologique seule.

langue fortement flexionnelle dont les paradigmes verbaux comportent jusqu'à 572 cases, dont 300 cases pour l'indicatif, qui sont remplies par 100 formes distinctes en raison de syncrétismes systématiques (Jacques *et al.*, 2012).

En khaling, une forme fléchie d'un verbe transitif marque à la fois le sujet et l'objet, d'une façon qui fait du système verbal de cette langue un système à direct-inverse (Silverstein, 1976 ; Zúñiga, 2006). Dans un système à direct-inverse, les personnes sont ordonnées selon une hiérarchie d'empathie (1, 2 > 3ANIM > 3INAN). On appelle formes *directes* les formes où le sujet est à une place supérieure à l'objet dans cette hiérarchie et formes *inverses* les autres formes. Dans un système à direct-inverse canonique (Walther *et al.*, 2014a), une forme directe et la forme inverse correspondante sont identiques, modulo un marqueur supplémentaire qui n'est porté que par la forme inverse [7]. Le khaling n'est pas un exemple de système à direct-inverse canonique, et s'en distingue par différents aspects à propos desquels nous renvoyons à (Jacques *et al.*, 2012 ; Walther *et al.*, 2014a).

Si l'on ignore le marqueur d'inverse *ʔi-*, le nombre de formes distinctes descend à 68, et certaines régularités peuvent être observées dans les paradigmes, régularités qui ne correspondent qu'en partie à des ensembles de cases réalisant des propriétés morphosyntaxiques homogènes. Un tel système peut donc être décrit au moyen de la notion de *trait morphomique*, des traits reflétant des propriétés structurales d'un système morphologique qui ne sont pas motivables par des considérations morphosyntaxiques, contrairement à des traits « classiques » tels que le genre ou le nombre (Aronoff, 1994, cf. également la section A.1.1). L'utilisation explicite de traits morphomiques est possible en PARSLI et Alexina PARSLI, où l'on peut les définir au moyen de règles de transfert (cf. section 2.2.1.6) [8]. Nous avons donc développé deux descriptions concurrentes, STD et MRPHM. La description STD repose sur un inventaire de traits morphosyntaxiques classique qui n'intègre pas de trait direct/inverse. La description MRPHM intègre le trait morphomique direct/inverse ainsi qu'un trait morphomique plus abstrait qui remplace les traits classiques afin de capturer les généralisations que nous avons mentionnées à l'instant. La description MRPHM est plus complexe au niveau de la spécification des traits, notamment parce qu'elle inclut des règles de transfert non-triviales qui permettent de définir les traits morphomiques à partir des traits morphosyntaxiques (cf. section 2.2.1.6), mais plus simple au niveau de la spécification des règles de réalisation, grâce aux généralisations supplémentaires qui sont captées. D'où la pertinence de notre mesure

7. Par exemple, en khaling, *ləpi* 'nous$_{\text{DU.INCL}}$ l'attrapons' et *ʔiləpi* 'il nous$_{\text{DU.INCL}}$ attrape' ne sont distingués que par le marqueur d'inverse qu'est le préfixe *ʔi*.

8. De telles règles permettent en effet de spécifier, pour un trait morphomique donné, chacune des valeurs possibles de ce trait et l'inventaire des cases du paradigme qui correspondent à chacune de ces valeurs.

quantitive pour comparer la complexité de ces deux descriptions, telle que mesurée par leur longueur de description selon la mesure décrite précédemment [9].

X \ Y	1SG	1DU.INCL	1DU.EXCL	1PL.INCL	1PL.EXCL	2.SG	2DU	2PL	3SG	3DU	3PL
1SG						loɔ̃m-nɛ	loɔ̃m-su	loɔ̃m-nu	lob-u	lob-u-su	lob-u-nu
1DU.INCL									lɵp-i		
1DU.EXCL						ʔi-loɔp	ʔi-lɵp-i	ʔi-loɔ̃m-ni	lɵp-u		
1PL.INCL									loɔp-ki		
1PL.EXCL						ʔi-loɔp	ʔi-lɵp-i	ʔi-loɔ̃m-ni	loɔp-kʌ		
2SG	ʔi-loɔ̃m-ŋʌ								ʔi-lɵ:b-ʉ	ʔi-lɵ:p-su	ʔi-lɵ:p-nu
2DU	ʔi-loɔ̃m-ŋʌ-su								ʔi-lɵp-i		
2PL	ʔi-loɔ̃m-ŋʌ-nu								ʔi-loɔ̃m-ni		
3SG	ʔi-loɔ̃m-ŋʌ								lɵ:b-ʉ		
3DU	ʔi-loɔ̃m-ŋʌ-su	ʔi-lɵp-i	ʔi-lɵp-u	ʔi-loɔp-ki	ʔi-loɔp-kʌ	ʔi-loɔp	ʔi-lɵp-i	ʔi-loɔ̃m-ni		lô:p-su	
3PL	ʔi-loɔ̃m-ŋʌ-nu										lô:p-nu

TABLEAU 4.2 – Paradigme positif non-passé du verbe khaling LOP ‘attraper’, representé à partir des traits morphosyntaxiques standard. La case située à l’intersection entre la ligne X et la colonne Y correspond à la forme notée traditionnellement X>Y, où X indique le nombre et le genre de l’agent et Y le nombre et le genre du patient.

X	X>1	1>X	B	C	D	E	F	G	H	I	J	K
2SG	loɔ̃m-ŋʌ	loɔ̃m-nɛ	lɵp-i	lɵp-u	loɔp-ki	loɔp-kʌ	loɔp	lɵp-i	loɔ̃m-ni	lɵ:b-ʉ	lô:p-su	lô:p-nu
2DU	loɔ̃m-ŋʌ-su	loɔ̃m-su										
2PL	loɔ̃m-ŋʌ-nu	loɔ̃m-nu										
3S	loɔ̃m-ŋʌ	lob-u										
3DU	loɔ̃m-ŋʌ-su	lob-u-su										
3PL	loɔ̃m-ŋʌ-nu	lob-u-nu										

TABLEAU 4.3 – Paradigme positif non-passé réduit du verbe khaling LOP ‘attraper’ grâce à l’utilisation de traits morphomiques et du trait direct/inverse. Quelques correspondances entre traits morphomiques et traits morphosyntaxiques standard : B : 1DU.INCL, F : AG=2SG, I : 2SG>3SG ou 3SG>3SG.

Nos deux descriptions ont été implémentées sur la base de 167 verbes produisant 50 100 formes, l’ensemble constitant le lexique KhaLex mentionné précédemment. L’application de notre mesure montre que la longueur de description de STD est de 15% supérieure à celle de MRPHM (la contribution du lexique morphologique est ignorée, puisqu’elle est constante d’une analyse à l’autre, à l’inverse de l’expérience précédente). Ces résultats, illustrés à la figure 4.2, montrent que l’utilisation du trait direct/inverse et de traits morphomiques permet une réduction significative de la compacité de la description, telle qu’estimée par notre mesure. En particulier, le coût additionnel lié à l’utilisation du trait supplémentaire direct/inverse et à la définition des traits morphomiques à partir des traits morphosytaxiques par les règles de transfert est plus que compensée par les généralisations qu’elle permet. Cela peut constituer une indication de la pertinence de ce

9. Une telle comparaison permet de quantifier l’intuition d’Aronoff (1994) qui, en définissant la notion de morphome, cherchait à capturer le niveau morphologique en tant que niveau d’organisation autonome doté de structures propres. L’identification et la représentation explicite de ces structures propres permet le développement des descriptions plus élégantes et plus économiques, comme mis en œuvre par exemple par Bonami et Boyé (2010).

type de traits pour la compréhension et la description de ce système morphologique de façon économique [10].

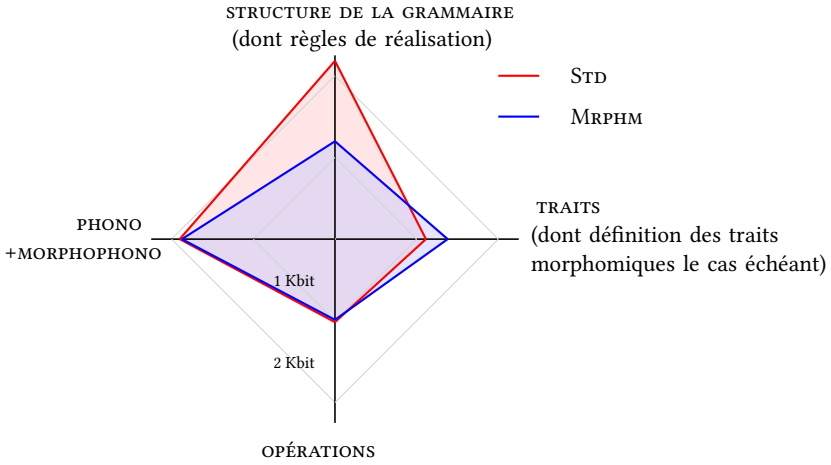


FIGURE 4.2 – Longueurs de description pour deux formalisations distinctes en Alexina PARSLI de la flexion verbale du khaling, réparties en composantes principales de la grammaire morphologique.

4.2 Complexité morphologique et prédictibilité entre cases

Comme nous l'avons indiqué précédemment, l'un des avantages des approches reposant sur des descriptions explicites (grammaire et lexique morphologique) est que l'on peut s'appuyer sur des formalismes morphologiques riches, à même de capturer un maximum de généralisations perçues comme potentiellement pertinentes. Il n'en reste pas moins que le développement manuel de descriptions est un travail coûteux et que ces descriptions peuvent être biaisées par le point de vue de leur auteur. De plus, la complexité d'un formalisme comme Alexina PARSLI rend difficile l'utilisation de la longueur de description dans un paradigme de type MDL : savoir comparer plusieurs descriptions est utile, mais il y aurait un grand pas à franchir avant de pouvoir produire la meilleure description possible dans le modèle choisi, c'est-à-dire la description la plus compacte. Enfin, l'ensemble de ces travaux repose sur l'hypothèse de la pertinence d'un tel modèle, et leurs résultats dépendent naturellement à la fois du modèle et de la façon de définir la mesure elle-même.

4.2.1 Le « Problème du Remplissage des Cases d'un Paradigme » (PCFP)

Les approches implicatives reposant sur l'étude de l'interprédictibilité entre cases des paradigmes, que nous évoquons plus en détail à la section A.4.2.2, constituent un autre

10. Il serait intéressant d'étudier l'influence de la taille du lexique sur ces résultats, par exemple en rejouant l'expérience avec les n lemmes les plus fréquents d'un corpus de la langue.

moyen de chercher à modéliser la complexité morphologique. La question sous-jacente est la recherche d'une forme de simplicité sous-jacente à l'apparente complexité des systèmes morphologiques, simplicité qui, intuitivement, devrait permettre à ces systèmes d'être appris et mis en œuvre par les locuteurs. Autrement dit, on ne cherche plus, dans cette section, à modéliser ou contrôler un certain arbitraire descriptif, mais à mieux comprendre les propriétés intrinsèques des systèmes morphologiques qui les rendent plus ou moins faciles à apprendre et à mettre en œuvre.

Comme nous le décrivons dans la section A.4.2.2, les travaux qui vont en ce sens s'appuient sur l'entropie informationnelle, et plus précisément sur l'entropie conditionnelle qui caractérise la tâche de prédiction d'une case sachant une autre. C'est de cette façon qu'Ackerman et ses collègues (Ackerman *et al.*, 2009 ; Malouf et Ackerman, 2010 ; Ackerman et Malouf, 2013) abordent le *problème du remplissage des cases d'un paradigme* (en anglais *Paradigm Cell Filling Problem*, ci-après *PCFP*), qu'(Ackerman *et al.*, 2009) formulent comme suit : « qu'est ce qui permet des inférences fiables sur les formes fléchies (et dérivées) de surface d'un item lexical ? [11] ». La reformulation de cette question au moyen de l'entropie conditionnelle consiste à mesurer la quantité d'information fournie par la connaissance d'une case donnée pour la prédiction d'une autre case : on mesure bien, dans le système morphologique étudié, le degré d'incertitude lié la façon de remplir une case si l'on en connaît une autre.

Les travaux d'Ackerman *et al.* (2009) partent d'une segmentation des formes fléchies en un radical et un suffixe. Ils en extraient alors un inventaire de classes flexionnelles au sein desquelles les correspondances entre cases, modélisées comme l'échange d'un suffixe par un autre, sont déterministes [12]. Ils calculent pour chaque classe flexionnelle et pour chaque couple de cases (c, c') l'entropie conditionnelle de c' sachant c. Ils définissent l'entropie conditionnelle globale de c' sachant c comme la moyenne des entropies conditionnelles correspondantes pour chacune des classes flexionnelles, chaque classe contribuant chacune de manière égale. Enfin, ils utilisent comme mesure de complexité du système, mesure appelée *entropie paradigmatique* (*paradigm entropy*), la moyenne des entropies conditionnelles de chacune des paires de cases (c, c').

Cette manière de procéder ne va pas sans poser des difficultés majeures, et notamment les suivantes, relevées par (Bonami *et al.*, 2011) et discutées plus en détail par (Bonami, 2014) :

— Les informations de fréquences ne sont pas prises en compte, chaque classe flexionnelle comptant autant que les autres, qu'elle contienne un seul verbe rare

11. What licenses reliable inferences about the inflected (and derived) surface forms of a lexical item?
12. On peut donc qualifier ces classes flexionnelles de *micro-classes* (Dressler, 2004 ; cf. également la notion de *plat* définie par Stump et Finkel)

ou de nombreux verbes fréquents [13]. On obtient ainsi un majorant des entropies que l'on calcule, ce qui peut suffire si l'on veut montrer que ces entropies sont plus basses que ce que l'on aurait attendu sous certaines hypothèses, mais qui ne fournit pas des résultats interprétables intrinsèquement.

— Les choix sur la façon de segmenter les formes en radical et suffixe, qui déterminent l'inventaire de classes flexionnelles, ont une influence sur les résultats [14]. Ceci est d'autant plus problématique que, pour ainsi dire, les locuteurs n'ont pas directement accès à des frontières entre morphes : comme discuté à la section A.4.2.2, il serait bien plus cohérent de ne pas partir de formes segmentées, et de ne s'appuyer que sur les formes brutes, seule méthodologie à même de s'affranchir en partie de ce type de biais et donc, notamment, de permettre d'envisager des comparaisons entre langues différentes [15].

— La (morpho)phonologie est intégrée aux données, de sorte que l'on ne mesure pas une complexité strictement morphologique. Par exemple, il peut être possible de lever certaines incertitudes captées par l'entropie conditionnelle par la seule connaissance de la forme du radical. Prendre en compte de telles fausses incertitudes constitue un artefact de la méthode.

— Le périmètre de l'étude est sujet à discussion : faut-il prendre en compte les verbes défectifs, et comment ? Par exemple, Bonami (2014, p. 127) fait le choix de ne pas les prendre en compte. Faut-il conserver les verbes irréguliers, susceptibles à eux seuls de modifier certaines valeurs de façon significative ? Répondre négativement à cette dernière question requiert naturellement une définition claire de ce qu'est un verbe irrégulier.

Nous avons présenté dans (Sagot, 2013a) diverses expériences sur le système verbal du français qui montrent à quel point la prise en compte de ces remarques peut jouer sur les valeurs que d'entropie paradigmatique que l'on obtient. Les résultats de ces expériences, ainsi que ceux publiés par Bonami *et al.* (2011), varient du simple au décuple, comme on peut le constater à la lecture de la table 4.4 : ils confirment l'importance qu'il y a à définir précisément la mesure que l'on retient et la façon dont on l'applique.

Nous avons également montré dans (Sagot, 2013a) que deux difficultés supplémentaires, peut-être plus fondamentales encore, affectent également la proposition initiale d'Acker-

13. Ce qui ne saurait refléter la réalité des données rencontrées par les néolocuteurs au cours processus d'apprentissage, qui est pourtant l'objet initial des études dans la lignée des travaux de Ackerman *et al.* (2009).

14. Bonami *et al.* (2011) citent ainsi l'exemple des verbes en *-ir* en français, pour lesquels une segmentation traditionnelle de l'infinitif (suffixe *-r* pour les verbes du deuxième groupe et *-ir* pour ceux du troisième groupe) rend déterministe le choix, pour certaines cases comme celles de l'imparfait, entre suffixes en *-iss-* (*finir* / *finissons*) et suffixes sans *-ss-* (*sortir* / *sortons*)

15. Nous avons qualifié les approches reposant sur l'entropie informationnelle au sein des paradigmes d'approches fondamentalement abstractives, ce qui est en contradiction avec le fait de partir de formes segmentées manuellement. Nous reviendrons ci-dessous sur ce point.

	Tous (fr)	« réguliers » (fr-reg)	1er groupe (fr1)	« réguliers » du 1er groupe (fr1-reg)
Segmentation manuelle (source : Le*fff*)				
Fréquences ignorées	0,70	0,42		0,40
Fréquences lexicales	0,10	0,07		0,03
Fréquences en corpus	0,19	0,17		0,02
Segmentation manuelle + désapplication des règles morphophonologiques (source : Le*fff*)				
Fréquences ignorées	0,64	0,28		0,25
Fréquences lexicales	0,08	0,06		0,03
Fréquences en corpus	0,16	0,13		0,02
Segmentation automatique par l'algorithme présenté en 3.1.2				
Fréquences ignorées	0,61	0,29	0,12	0,17
Fréquences lexicales	0,14	0,10	0,08	0,08
Fréquences en corpus	0,27	0,20	0,08	0,08
Segmentation automatique par l'algorithme de Bonami *et al.* (2011)				
Fréquences ignorées	0,68			
Fréquences lexicales	0,74			

Tableau 4.4 – Comparaison entre les entropies paradigmatiques obtenues pour le système verbal du français selon les choix retenus relativement aux informations de fréquence, le mode de segmentation des formes et à la prise en compte des verbes non « réguliers », c'est-à-dire appartenant à des classes flexionnelles rares (moins de 10 membres). Prendre en compte la « fréquence lexicale » consiste à pondérer chaque classe flexionnelle par le nombre d'entrées lexicales qui l'utilisent. Prendre en compte la « fréquence en corpus » consiste à pondérer chaque classe flexionnelle par la somme des fréquences en corpus des lemmes qui l'utilisent, le corpus utilisé étant le Corpus Arboré de Paris 7.

man et de ses collègues, mais également certaines propositions ultérieures qui en sont des améliorations significatives, notamment par rapport à certain des points précédents (Bonami *et al.*, 2011). Ces deux difficultés supplémentaires sont les suivantes :

— L'approche implicative, bien que s'appuyant sur des calculs réalisés directement sur les paradigmes, est tout aussi dépendante du modèle formel sous-jacent que les approches reposant sur des descriptions, le modèle formel dont il est question ici étant celui qui permet la formalisation du rapport formel entre deux formes. Considérons par exemple un système morphologique où la forme remplissant une certaine case c' est systématiquement la réduplication de celle remplissant une autre case c. Si la réduplication fait partie des opérations disponibles pour modéliser les relations entre formes, l'entropie de c' sachant c sera nulle. Si, comme dans le travail d'Ackerman *et al.* (2009), la seule opération disponible est le remplacement

d'un suffixe (éventuellement vide) par un autre (éventuellement vide), l'entropie conditionnelle de c' sachant c sera très élevée [16].

— La taille des paradigmes est ignorée. Considérons par exemple deux systèmes morphologiques : le *système complet* les verbes du premier groupe en français et le *système réduit* que l'on peut en extraire en ne conservant que deux cases, l'infinitif et la première personne du pluriel de l'indicatif présent. Si l'on applique les propositions d'Ackerman *et al.* (2009), on obtient une entropie paradigmatique supérieure pour le système réduit par rapport au système complet [17]. La raison en est simple : l'entropie paradigmatique est la *moyenne* des entropies conditionnelles de toutes les paires de cases ; or il existe dans le système complet de nombreuses paires de cases qui sont fortement voire totalement interprédictibles, et ces paires de cases font baisser la moyenne ; ce n'est pas le cas pour le système réduit, composé de deux cases entre lesquelles l'interprédictibilité est plus limitée. À cet égard, les conclusions de Bonami *et al.* (2011) et celles de Bonami (2014, section 3.3) sur la comparaison du système verbal du mauricien avec celui du français nous semblent discutables. Le système verbal du mauricien, structurellement similaire au système réduit que nous venons de considérer, est formé de deux cases. L'application de la notion d'*entropie implicative*, version améliorée de la mesure proposée par (Ackerman *et al.*, 2009), conduit au résultat suivant : « l'entropie implicative est incontestablement plus élevée pour le mauricien que pour le français ». De ce constat, il est tiré la conclusion suivante : « Si on admet, à la suite d'Ackerman et Malouf (2013), que le PCFP est une composante importante de la complexité d'un système morphologique, la conclusion qui s'impose est que le système flexionnel du mauricien est plus complexe que celui du français sur ce point particulier ». Toute la question est donc de savoir si « ce point particuler » est pertinent [18]. Mais il est douteux que l'on puisse dire à propos de notre système réduit par rapport à notre système complet ce que Bonami (2014) écrit à propos du mauricien par rapport au français : « là où une langue créole a de la morphologie flexionnelle substantielle, il semble qu'elle puisse donner lieu aux phénomènes d'opacité familiers des langues à morphologie riche, *et dans des proportions qui vont au-delà de ce qu'on peut observer dans sa langue lexificatrice* » (c'est nous qui soulignons).

Répondre convenablement à la première de ces difficultés est aujourd'hui encore un problème ouvert. Il peut être approché sous l'angle de la modélisation des relations entre cases ou sous celui de l'inférence de classes de lemmes au comportement flexionnel

16. Cet argument est repris par Baerman *et al.* (2015), où l'exemple proposé dans (Sagot, 2013a) pour illustrer ce phénomène est repris et discuté.

17. Et ce, que les informations de fréquence (dans le lexique ou en corpus) soient prises en compte ou non.

18. Notamment pour des travaux qui, comme pour Ackerman *et al.* (2009), sont motivés par des problématiques d'ordre cognitif.

similaire, étant entendu que la notion de similarité dont il est question dépend directement de la façon dont on envisage la modélisation des relations entre cases. Plusieurs approches non-supervisées ont été proposées à cette fin, telles que l'utilisation de mesures globales de similarité entre paradigmes, par exemple au moyen d'une distance de compression (Brown et Evans, 2012). Si une telle approche est fondée théoriquement, elle repose en pratique sur l'implémentation de l'algorithme de compression utilisée [19], et résulte donc *in fine* d'une façon de modéliser la structure des formes et des paradigmes, même si cette modélisation est très générique. En réalité, ces questions se ramènent à celle de la modélisation quantitative de la notion de classe flexionnelle, sur laquelle nous reviendrons à la section 4.3.

Une autre proposition consiste, pour chaque paradigme, à identifier dans les formes qui le composent une partie commune maximale et un jeu d'affixes, à l'image de ce que nous avons proposé en section 3.1.2, puis à utiliser une mesure de similarité faisant usage de cette segmentation (Lee et Goldsmith, 2013). Une dernière approche consiste à comparer, d'un lemme à l'autre, les motifs d'alternance caractérisant toutes les paires de formes de leur paradigme (Bonami, 2014). Mais ces deux dernières approches dépendent toujours de façon massive de la façon dont on modélise les relations entre formes. Nous reviendrons sur cette problématique dans les prochaines sections de ce chapitre.

4.2.2 Complexité Paradigmatique Globale Minimale

La seconde des difficultés fondamentales évoquées ci-dessus peut être surmontée si l'on revient à la notion même de complexité morphologique, afin de comprendre ce que l'on peut chercher à capter par une telle notion. Tout d'abord, il nous semble intuitif de considérer que la complexité morphologique (flexionnelle) doive capter le degré d'irrégularité de chacun des paradigmes, cette irrégularité ne pouvant être définie que par rapport à l'ensemble de ces paradigmes. Elle est mesurée à partir des motifs d'alternances entre formes, le plus souvent *via* l'entropie d'une case conditionnellement à une autre. La proposition d'Ackerman *et al.* (2009), qui rentre dans ce cadre, souffre de ce que cette irrégularité est mesurée comme une moyenne des entropies conditionnelles entre cases du paradigme, avec les conséquences indiquées ci-dessus.

Mais une façon plus naturelle d'appréhender ce concept d'irrégularité consiste à considérer la quantité d'information nécessaire à la construction du paradigme complet d'un lemme étant donnés les alternances formelles entre cases et les entropies conditionnelles associées : dans le cas d'un système parfaitement régulier, toutes les entropies conditionnelles sont nulles, et l'on peut produire l'ensemble des formes à partir de l'une d'entre elles sans aucune information complémentaire. L'irrégularité est donc

19. Lequel algorithme est du reste trop complexe pour être véritablement adapté à la tâche pour laquelle il est mis en œuvre par les auteurs.

nulle. Un système flexionnel totalement aléatoire ne permettrait pas la construction d'un paradigme particulier par un moyen plus compact que la donnée explicite (extensionnelle) dudit paradigme. L'irrégularité est alors maximale. Comme évoqué à l'instant, faisons pour l'instant l'hypothèse simplificatrice que l'input du processus de construction des paradigmes est une forme unique.

Mettre en pratique l'intuition que nous venons d'énoncer consiste donc à déterminer quelle est la case d'input permettant de reconstituer un paradigme complet à partir de la forme remplissant cette case, de telle façon que la quantité d'information nécessaire pour ce faire soit minimale. La quantité d'information ainsi obtenue est ce que nous avons appelé la Complexité Paradigmatique Globale Minimale (*Minimum Overall Paradigm Complexity*, MOPC ; Sagot, 2013a). Calculer la MOPC d'un système flexionnel consiste donc à choisir une case d'input et un arbre de dérivation reliant cette case d'input à toutes les autres cases, de façon directe ou indirecte, de telle sorte que cet arbre minimise la *somme* (et non la *moyenne*) des entropies conditionnelles correspondant à chacun de ses arcs. Ainsi, on mesure la quantité d'information nécessaire pour spécifier comment construire, à partir de la case d'input optimale, les formes remplissant les cases qui sont ses fils dans cet arbre, puis celles remplissant les cases qui sont les fils de ces dernières, et ainsi de suite jusqu'à avoir rempli tout le paradigme. Une telle mesure est inchangée si l'on rajoute des cases dont les formes sont calculables de façon déterministe à partir d'autres cases. De plus, elle ne peut que décroître (ou rester constante) si l'on enlève des cases. Ces deux propriétés, qui semblent pourtant intuitivement souhaitables, ne sont pas vérifiées par l'entropie paradigmatique proposée par Ackerman *et al.* (2009) [20].

Le calcul pratique de la MOPC revient donc à extraire l'arbre couvrant de poids minimal (*minimum spanning tree*) dans le graphe complet reliant toute case à une autre avec pour poids l'entropie conditionnelle de la case cible sachant la case source, puis à calculer la somme des poids des arcs de cet arbre [21]. Notre implémentation utilise l'algorithme classique de Chu–Liu-Edmonds (Chu et Liu, 1965 ; Edmonds, 1967) pour extraire cet arbre.

Reste à calculer l'entropie conditionnelle de toutes les cases sachant toutes les autres, comme pour le calcul de l'entropie paradigmatique. Ici encore, un tel calcul repose nécessairement sur une segmentation qui peut être manuelle ou automatique, dichotomie discutée précédemment. La table 4.5 donne les résultats du calcul de la MOPC à partir d'une segmentation automatique identique à celle utilisée pour obtenir les entropies paradigmatiques données à la table 4.4. La figure 4.3 donne l'arbre couvrant minimal

20. On notera toutefois que la première de ces deux propriétés a pour résultat qu'une certaine forme de complexité, à savoir la structure-même des paradigmes (leur taille par exemple) n'est pas du tout prise en compte par la MOPC. Il s'agit d'un exemple de plus du fait qu'il est difficile de définir *la* complexité morphologique, et que la MOPC, comme d'autres mesures, capturent chacune *une* forme de complexité morphologique.

21. Puisque nous manipulons donc un graphe orienté, ce que l'on recherche n'est pas à proprement parler l'*arbre* couvrant de poids minimal, mais l'*arborescence* couvrante de poids minimal. Nous utiliserons néanmoins le terme d'*arbre* par abus de langage.

	Tous (fr)	« réguliers » (fr-reg)	1er groupe (fr1)	« réguliers » du 1er groupe (fr1-reg)
Segmentation automatique par l'algorithme présenté en 3.1.2				
Fréquences ignorées	2.43 inf	0.85 cond.pst =ind.fut	0.05 cond.pst =ind.fut	0.13 cond.pst =ind.fut
Fréquences lexicales	0.52 cond.pst.3s	0.26 ind.pst.pfv.s =sbjv.pst	0.13 ptcp.pst =…	0.13 ptcp.pst =…
Fréquences en corpus	0.80 inf	0.58 ind.prs.1s	0.08 ptcp.pst =…	0.08 ptcp.pst =…

Tableau 4.5 – Comparaison entre les MOPC obtenues pour le système verbal du français selon les choix retenus relativement aux informations de fréquence, le mode de segmentation des formes et à la prise en compte des verbes non « réguliers », c'est-à-dire appartenant à des classes flexionnelles rares (moins de 10 membres). La case ou tout ou partie des cases d'input obtenue(s) est/sont indiquée(s) dans chaque cas (la notation ptcp.pst =… est à lire comme ptcp.pst =ind.prs.1p =ind.pst.pfv.s =sbjv.pst).

correspondant à la MOPC de 0,80 obtenue avec prise en compte des fréquences en corpus et sans filtrage sur les comportements flexionnels rares. Sans surprise, c'est l'infinitif qui est le point de départ optimal pour dériver les autres cases.

Si, avec la MOPC, la dépendance à la taille du paradigme est réglée par l'extraction de l'arbre couvrant minimal, les autres difficultés citées précédement demeurent : difficultés liées à la prise en compte des informations fréquentielles, rôle peu clair de la morphophonologie, dépendance au modèle formel permettant d'exprimer les transformations d'une forme à une autre, et notamment, si le modèle utilisé est concaténatif, arbitraire de segmentation (qu'elle soit manuelle ou automatique). De plus, concernant spécifiquement la MOPC telle que définie ici, nous avons fait deux approximations : la première consiste à ne partir que d'une seule forme pour inférer les autres, la seconde réside dans ceci que tous les lemmes sont traités comme un ensemble unique dans lequel on mesure des entropies conditionnelles : on fait comme s'il n'y avait qu'une seule (macro)classe flexionnelle, quand bien même sont naturellement distingués les comportements flexionnels distincts (les micro-classes). Nous reviendrons sous peu sur ce point et sur ces concepts de macro-classe et de micro-classe.

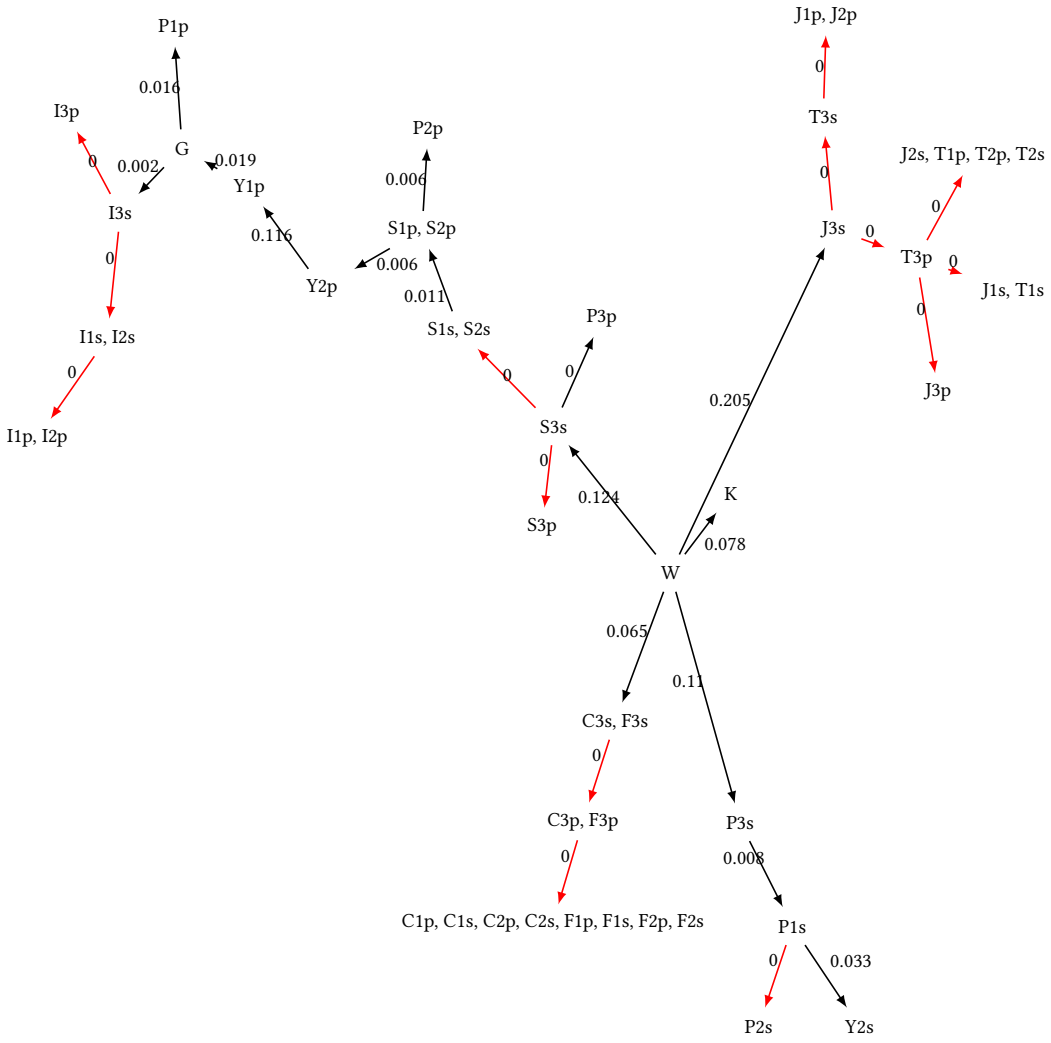


FIGURE 4.3 – Arbre couvrant minimal correspondant à la MOPC de 0,80 obtenue pour le système verbal du français avec prise en compte des fréquences en corpus et sans filtrage sur les comportements flexionnels rares. Les cases sont identifiées par leur étiquette morphologique L*ffff*, reprise de l'inventaire du projet MULTEXT (cf. section 2.1, note 9 pour plus de détails et des exemples).

4.3 Inférence endogène d'une hiérarchie de classes flexionnelles [22]

Nous avons jusqu'ici proposé deux approches fondamentalement différentes pour quantifier la complexité morphologique : (i) une approche constructive reposant sur la notion de longueur de description pour comparer des descriptions morphologiques manuelles concurrentes en termes d'économie descriptive et (ii) une approche abstractive reposant sur l'entropie conditionnelle des alternances formelles entre cases des paradigmes et sur la notion d'arbre couvrant minimal pour calculer une approximation de la quantité d'information nécessaire à la connaissance d'un paradigme complet à partir de la forme globalement maximalement prédictive, approximation que nous avons appelée la MOPC.

Nous avons déjà discuté des limites inhérentes à chacune de ces approches. Revenons sur deux d'entre elles. D'une part, l'approche reposant sur la longueur de description souffre de l'arbitraire lié au fait même qu'elle repose sur des descriptions morphologiques et non sur les paradigmes eux-mêmes, difficulté que les approches abstratives cherchent à éviter. D'autre part, la MOPC, pour simplifier, ignore la notion de classe flexionnelle, contrairement à la première approche qui s'appuie sur des descriptions morphologiques qui en font un usage systématique, quoique soumis comme nous venons de le rappeler à l'arbitraire du morphologue qui construit les descriptions [23]. Ainsi, elle ne prend pas en compte la diversité structurée des comportements flexionnels.

La notion de classe flexionnelle est une notion intuitive très généralement répandue, indépendamment des approches et des partis pris, à la fois chez de nombreux grammairiens anciens, dans l'enseignement des langues et en morphologie formelle (cf. par exemple Blevins, 2016, sec. 1.3.2). Mais cette notion reste difficile à formaliser, notamment dans un cadre abstractif. Partons donc d'une définition purement intuitive : une classe flexionnelle peut être définie comme un ensemble de lemmes qui se fléchissent de façon similaire — nous parlerons de *similarité inter-paradigmatique* [24]. Ce point de départ va nous permettre de tenter de modéliser les systèmes flexionnels d'une façon qui, tout en reposant sur une approche purement abstractive et sera donc non-supervisée [25],

22. Le contenu de cette section résulte d'un travail collaboratif avec Olivier Bonami et Sarah/Sacha Beniamine dans le cadre de l'opération MORPHO1 de l'axe 2 « Grammaire expérimentale » du LabEx EFL. Il a notamment fait l'objet de deux communications orales (Beniamine *et al.*, 2015 ; Beniamine et Sagot, 2015) et d'un article (Beniamine *et al.*, 2018).

23. Dans notre cas, il s'agit en fait de schèmes flexionnels tels que définis par ᖴᗩᖇᔕᒪᓀ et Alexinaᖴᗩᖇᔕᒪᓀ, mais l'idée sous-jacente est bien la même. Voir la note suivante.

24. Les réflexions poursuivies dans cette section reposent sur une intuition de ce qu'est une classe flexionnelle qui est moins approfondie, au moins à certains égards, que ce qui a conduit en ᖴᗩᖇᔕᒪᓀ aux notions de zone flexionnelle et de schème flexionnel. Il y a donc ici, en partie, simplification du problème posé. Il n'en reste pas moins qu'il sera nécessaire, à l'avenir, de poursuivre ce travail de façon à prendre en compte les différents phénomènes non-canoniques définis et modélisés au chapitre 2. Nous y reviendrons plus bas.

25. Nous employons ici le terme de *non-supervisé* pour indiquer qu'aucune information et aucun exemple relevant de la notion de classe flexionnelle n'est donnée en entrée au système. En revanche, les données

reposera crucialement sur la notion de classe flexionnelle et nous permettra de faire intervenir la notion d'économie descriptive au travers d'une mesure de longueur de description.

Les descriptions des systèmes de classes flexionnelles prennent, dans la littérature, des formes variées qui vont d'inventaires plats à des hiérarchies à héritage multiple, comme par exemple en *Network Morphology* (Corbett et Fraser, 1993 ; Brown et Hippisley, 2012). L'inférence non-supervisée d'un inventaire plat de classes flexionnelles a été évoquée de façon rudimentaire à la section 3.1.2. Naturellement, des travaux plus approfondis ont été publiés pour l'extraction automatique de systèmes plus complexes de classes flexionnels à partir de lexiques morphologiques extensionnels, y compris à des fins typologiques et théoriques (Brown et Evans, 2012 ; Lee et Goldsmith, 2013 ; Bonami, 2014). Tous reposent sur une modélisation de la notion de similarité entre comportements flexionnels, que nous appellerons *similarité interparadigmatique*. Ils diffèrent en revanche sur la façon de modéliser et de mesurer cette similarité interparadigmatique, et notamment sur les stratégies de segmentation mises en œuvre à cette fin pour identifier des parties communes et des parties variables entre formes. Plus généralement, l'inférence non-supervisée d'un système de classes flexionnelles dépend *a priori* des réponses que l'on apporte aux questions fondamentales suivantes : (i) quelles peuvent être les structures possibles pour un système de classes flexionnelles ? (ii) sur quels types de généralisations (elles-mêmes résultant généralement de techniques de segmentation) s'appuie-t-on pour mesurer la similarité interparadigmatique entre lemmes ? (iii) comment quantifier cette similarité interparadigmatique et comment en déduire une mesure de scorage de classes flexionnelles candidates, mesure dont on aimerait qu'elle minimise une certaine forme de complexité ? Nous allons passer ces questions en revue successivement, ce qui nous permettra de motiver les choix que nous avons effectués lors de la conception de notre algorithme d'extraction non-supervisée de hiérarchies de classes flexionnelles. Nous verrons que la notion de minimisation de la longueur de description, proposée initialement dans le cadre constructif de la section 4.1, est mise en œuvre dans un contexte abstractif, qui fait ainsi suite à certaines des réflexions de la section 4.2.

4.3.1 Micro-classes et macro-classes

Comme nous l'avons indiqué, la notion intuitive de classe flexionnelle repose sur la similarité interparadigmatique. Sa modélisation repose donc sur celle de cette similarité. La façon la plus générale de procéder est de disposer d'une définition de la similarité paradigmatique qui soit scalaire : deux comportements flexionnels peuvent être plus ou moins similaires. La façon la plus courante de modéliser cette gradualité est de définir des

d'entrée utilisées dans cette section sont un lexique morphologique extensionnel, lexique qui inclut nécessairement les paradigmes complets des lemmes qu'il contient (contrairement à un lexique extrait directement d'un corpus, par exemple, où rares sont les lemmes dont toutes les formes fléchies sont attestées).

hiérarchies de comportements flexionnels (Corbett et Fraser, 1993 ; Dressler et Thornton, 1996 ; Kilani-Schoch et Dressler, 2005 ; Haspelmath et Sims, 2010 ; Brown et Hippisley, 2012 ; Lee et Goldsmith, 2013). Reste à comprendre comment on peut extraire de façon non-supervisée une structure aussi complexe qu'une hiérarchie de classes flexionnelles — et surtout, selon quels principes fondamentaux.

Corbett (2009), mettant ainsi en œuvre les principes de la typologie canonique, propose de pousser à l'extrême l'intuition de départ pour définir une notion de classe flexionnelle canonique qui puisse servir de point extrémal dans la description des classes flexionnelles que l'on est amené à manipuler. Dans un système de classes flexionnelles canoniques, chaque classe est maximalement homogène en interne et maximalement hétérogène en externe, c'est-à-dire qu'il n'y a aucune généralisation commune possible entre deux classes flexionnelles distinctes.

On peut donc imaginer deux stratégies distinctes pour construire un système de classes flexionnelles à partir d'un système morphologique donné, selon que l'on part de l'un ou l'autre des deux critères définissant la notion de classe flexionnelle canonique. La première consiste à utiliser une version stricte de la notion de similarité interparadigmatique, en la ramenant à une identité stricte entre comportements flexionnels. On peut qualifier de *micro-classes* les classes flexionnelles ainsi obtenues, c'est-à-dire les ensembles de lemmes se fléchissant de façon identique. Cette notion correspond à la notion de *plat* définie par Stump et Finkel (2013). Ces micro-classes sont homogènes en interne, mais elles peuvent partager un certain nombre de points communs, s'éloignant ainsi de la définition canonique indiquée ci-dessus. On peut les assimiler aux feuilles d'une hiérarchie complète de classes flexionnelles. Une telle définition reste dépendante de ce que l'on qualifie de comportements flexionnels identiques, c'est-à-dire qu'elle dépend de la façon dont on modélise les correspondances entre formes, et donc de la stratégie de segmentation utilisée, qui dépend elle-même, entre autres, de la place que l'on confère à la morphophonologie voire à la phonologie.

Une autre stratégie consiste donc à se doter d'un système de classes flexionnelles qui garantisse leur hétérogénéité externe, c'est-à-dire l'impossibilité qu'il y a à formuler des généralisations pertinentes communes à plusieurs classes. On obtient alors des ensembles de lemmes plus conséquents, que nous qualifierons de *macro-classes*. Chaque macro-classe contient des lemmes dont les paradigmes flexionnels ont au moins certains points en communs, mais cela ne saurait nécessairement impliquer qu'ils ont tous des comportements flexionnels identiques : cette fois-ci, c'est l'homogénéité en interne qui n'est plus assurée. Une description complète doit alors modéliser les variations dans le comportement flexionnel des lemmes au sein d'une même macro-classe, comme le font du reste la plupart des descriptions morphologiques traditionnelles. On peut imaginer identifier ces macro-classes à un ensemble particulier de nœuds au sein d'une hiérarchie

complète de classes flexionnelles. Cependant, il reste à comprendre comment formaliser la notion d'hétérogénéité maximale, sans quoi il sera impossible d'extraire automatiquement des macro-classes. Il est souvent fait usage, dans les descriptions traditionnelles, de critères variés qui, souvent, ne sont pas de nature morphologique, mais notamment sémantiques (animé/inanimé), phonologiques (verbes en *-cer*) ou morphosyntaxiques (genre).

Illustrons brièvement la distinction entre macro-classe et micro-classe en français. Intuitivement, et dans la quasi-totalité des descriptions du système verbal du français, il est fait usage d'une macro-classe intitulée « verbes du premier groupe » : les verbes qui appartiennent à ce groupe ont de nombreux points communs. Toutefois, ces différences correspondent à un ensemble de plusieurs micro-classes, comme par exemple celle des verbes de type JETER(*je jette*, *nous jetons* ...), celle des verbes de type HALETER(*je halète*, *nous haletons* ...), ou encore celle des verbes les plus « prototypiques » de ce premier groupe, celle du verbe AIMER(*j'aime*, *nous aimons* ...).

4.3.2 Inférence automatique d'une hiérarchie de classes flexionnelles à partir d'un lexique flexionnel extensionnel

Extraire automatiquement des micro-classes et des macro-classes nécessite donc de poser un certain nombre d'hypothèses et de principes de départ. Nous avons tout d'abord comparé deux façons distinctes de définir la notion de similarité elle-même (et donc, entre autres, les micro-classes), à partir de deux stratégies distinctes de segmentation, une stratégie locale et une stratégie globale, que nous avons développées et comparées. La stratégie globale consiste à analyser simultanément toutes les formes d'un paradigme et à en extraire les parties variables et les parties constantes. Si l'on ne s'autorise qu'une stratégie affixale, on est dans le cas décrit à la section 3.1.2. Pour le travail décrit ici, nous avons développé un algorithme plus complexe, qui permet l'identification de séquences alternant segments constants et segments variables, ce qui permet de capter les alternances vocaliques et autres phénomènes non-concaténatifs mais néanmoins segmentaux. La stratégie de segmentation locale fait usage du même algorithme, mais l'applique paire de cases par paire de cases. La table 4.6 permet de comparer le résultat de ces deux approches sur trois sous-paradigmes verbaux du portugais. On notera que les approches constructives de la morphologie, qui s'appuient sur la notion d'exposant [26], reposent de façon quasi-systématique sur des stratégies de segmentation globales, notamment pour minimiser le nombre de radicaux par lemme. À l'inverse, les approches abstractives restent souvent au niveau de la comparaison entre cases, y compris pour la segmentation, et utilisent donc la stratégie locale. Toutefois, il n'y a pas de nécessité dans

26. Un exposant, qui relie des caractéristiques de surface d'une forme fléchie à des propriétés morphosyntaxiques exprimées par cette forme, n'est pas nécessairement un morphe contigu : il peut être discontinu voire suprasegmental.

ces corrélations, et nous avons comparé les deux stratégies bien que notre approche soit ici abstractive. Du reste, on peut définir, comme par exemple le font Montermini et Bonami (2013), des stratégies intermédiaires, semi-locales, que l'on pourrait mettre en rapport avec la notion traditionnelle de partie principale ou encore avec la notion de zone flexionnelle définie par PARSLI et utilisée par AlexinaPARSLI (cf. chapitre 2). La figure 4.4 illustre les relations implicatives entre les cases du paradigme d'un adjectif espagnol selon que l'on s'appuie sur la stratégie de segmentation locale ou globale.

	PER.INF.1PL	GER	IND.PRS.3PL
ABANDONAR 'abandon'	ɐbɐ̃dunaɾmuʃ	ɐbɐ̃dunɐ̃du	ɐbɐ̃donɐ̃ũ
REABRIR 'reopen'	riɐbɾiɾmuʃ	riɐbɾĩdu	riabɾɐ̃ĩ
VOAR 'fly'	vuaɾmuʃ	vuɐ̃du	voɐ̃ũ

Paradigmes partiels

	PER.INF.1PL ⇌ IND.PRS.3PL	GER ⇌ IND.PRS.3PL	PER.INF.1PL⇌GER
ABANDONAR	_u_aɾmuʃ ⇌ _o_ɐ̃ũ	_u_du ⇌ _o_ũ	_aɾmuʃ ⇌ _ɐ̃du
REABRIR	_ɐ_iɾmuʃ ⇌ _a_ɐ̃ĩ	_ɐ_ĩdu ⇌ a_ɐ̃ĩ	_iɾmuʃ ⇌ _ĩdu
VOAR	_uaɾmuʃ ⇌ _oɐ̃ũ	_u_du ⇌ _o_ũ	_aɾmuʃ ⇌ _ɐ̃du

Segmentation locale

	PER.INF.1PL	GER	IND.PRS.3PL
ABANDONAR	_u_aɾmuʃ	_u_ɐ̃du	_o_ɐ̃ũ
REABRIR	_ɐ_iɾmuʃ	_ɐ_ĩdu	_a_ɐ̃ĩ
VOAR	_uaɾmuʃ	_uɐ̃du	_oɐ̃ũ

Segmentation globale

TABLEAU 4.6 – Extrait de paradigmes verbaux du portugais et patrons induits respectivement par les stratégies de segmentation locale et globale.

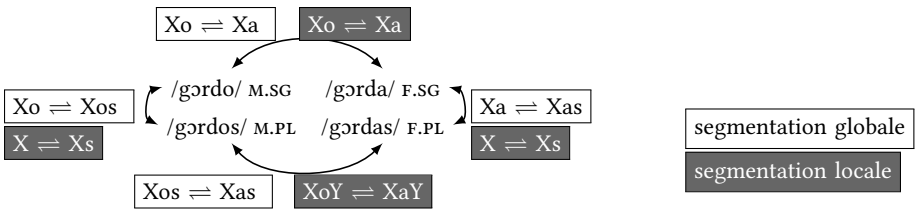


FIGURE 4.4 – Relations implicatives entre les formes de l'adjectif espagnol GORDO 'gros, gras' selon que l'on emploie la stratégie de segmentation locale ou globale.

Un autre choix déterminant est celui de la façon dont nous allons exploiter ces segmentations pour définir quantitativement la similarité interparadigmatique qui nous permettra tout d'abord de construire les micro-classes puis de chercher à les combiner entre elles pour former des classes de plus en plus importantes. Comme indiqué plus haut, il serait souhaitable de savoir arrêter ce processus de combinaison à un niveau qui

corresponde au mieux à la notion intuitive de macro-classe. C'est entre autres pour cette raison que les techniques les plus usuelles, qui s'appuient sur des distances (Brown et Evans, 2012 ; Bonami, 2014) ou sur l'entropie paradigmatique d'Ackerman et Malouf (2013) ne peuvent pleinement convenir : certaines de ces techniques ont l'avantage d'associer à deux lemmes appartenant à une même micro-classe une distance nulle, mais aucune ne prend en compte la cohérence ou la structure du système de classes flexionnelles dans son ensemble et, en conséquence, ne peut permettre l'identification des macro-classes. Nous avons donc choisi de convoquer à nouveau la notion de longueur de description utilisée à la section 4.1 afin de formaliser le principe définitoire selon lequel une macro-classe ne partage aucune généralisation utile avec des ensembles de lemmes qui n'en font pas partie. En effet, un système donné de classes flexionnelles peut servir de base à une description intensionnelle du système morphologique, description dont on peut estimer la longueur de description sans pour autant avoir besoin de la spécifier explicitement, par le calcul direct des valeurs d'entropie concernées [27]. Dans notre cas, contrairement à ce que nous avons présenté à la section 4.1, cette description intensionnelle est de nature fondamentalement abstractive : elle contient les informations nécessaires permettant de reconstituer les patrons d'alternance corrects [28].

L'algorithme est alors simple, et consiste en un *clustering* ascendant : on part d'un ensemble de classes flexionnelles formé des micro-classes, lesquelles rassemblent des lemmes au comportement flexionnel identique. Le système de classes flexionnelles ainsi obtenu correspond à une description morphologique dont on calcule la longueur de description. On calcule alors, pour chaque paire de classes, la longueur de description morphologique que l'on obtiendrait si on procédait à la fusion des deux classes concernées. L'idée est que fusionner deux classes partageant suffisamment de généralisations conduira à une description plus courte, malgré les informations qu'il faudra y rajouter pour spécifier les différences entre comportements flexionnels au sein de la classe fusionnée. À l'inverse, si les généralisations que l'on peut factoriser entre les descriptions de deux classes ne suffisent pas à compenser l'augmentation de la longueur de description induite par la description de ces différences, alors la fusion n'est pas légitime. Ainsi, on peut fusionner de proche en proche des classes de façon ascendante, jusqu'à ce qu'aucune fusion ne permette plus de faire diminuer la longueur de description. Les racines des arbres de classes flexionnelles ainsi obtenus peuvent alors être considérées comme des macro-classes : s'il n'y a plus aucune fusion qui permette de faire diminuer la longueur de description, c'est qu'il n'y a plus aucune généralisation utile à identifier. On peut néanmoins continuer à

27. Nous ne rentrerons pas ici dans le calcul détaillé de cette longueur de description. Nous renvoyons pour cela notamment à (Beniamine *et al.*, 2015, 2018).

28. Cette proposition a des points communs avec le travail de Lee et Goldsmith (2013). Cependant, ces derniers partent d'hypothèses et de données de nature différente, utilisent une stratégie de segmentation bien plus naïve et font usage d'une manière peu convaincante d'instancier formellement dans leur cas particulier la notion de longueur de description.

appliquer l'algorithme pour procéder aux regroupements qui conduisent à la plus petite dégradation de la longueur de description minimale, jusqu'à obtenir un arbre unique.

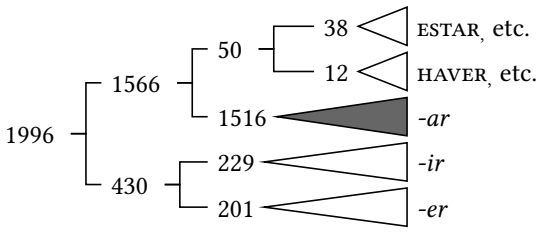
4.3.3 Expériences sur les systèmes verbaux du français et du portugais

Nous avons implémenté et appliqué cet algorithme, en comparant l'utilisation de la stratégie de segmentation globale et locale, sur des lexiques phonémiques verbaux du français (Bonami *et al.*, 2014) et du portugais européen (Veiga *et al.*, 2013). Dans les deux cas, l'utilisation de la stratégie de segmentation globale conduit à la construction d'un système de classes flexionnelles qui est très similaire au système traditionnel, comme l'illustre la figure 4.5 sur le portugais. À l'inverse, la stratégie globale produit un nombre plus élevé de macro-classes disparates. On peut y voir une conséquence du fait que la stratégie de segmentation globale fait inévitablement moins de généralisations pertinentes que la stratégie locale. Ce phénomène illustré par la table 4.6 où l'on peut voir que la stratégie globale ne permet pas d'extraire la moindre généralisation commune entre les lemmes ABANDONAR et VOAR, alors que la stratégie locale produit, dans deux cas sur trois, des patrons d'alternance identiques pour ces deux lemmes.

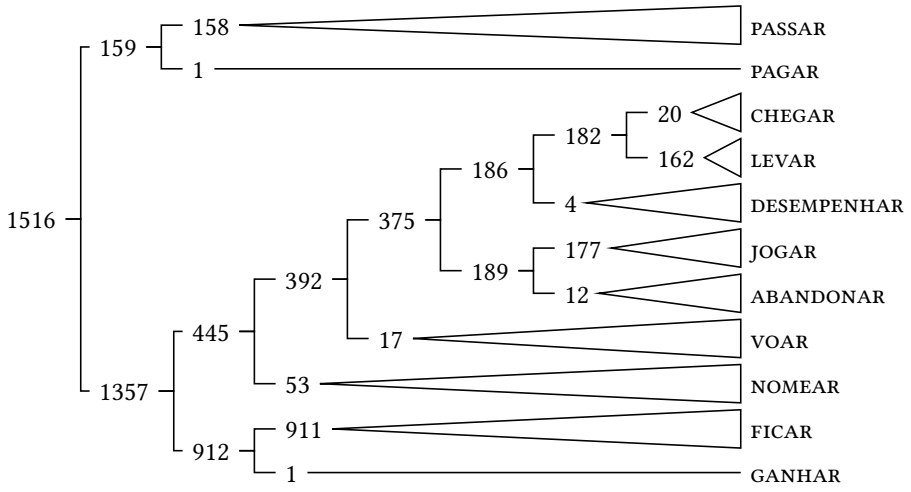
4.3.4 Discussion

Au vu de ces résultats, on pourrait vouloir conclure à une meilleure adéquation de la stratégie de segmentation locale par rapport à la stratégie globale. Toutefois, de nombreuses questions restent en suspens, dont la moindre n'est pas de comprendre tout simplement pourquoi l'approche proposée ici donne-t-elle, avec la stratégie de segmentation locale, des résultats si proches des classifications traditionnelles — les deux langues traitées disposant d'une tradition grammaticale fournissant de telles classifications. Après tout, les longueurs de description calculées et comparées ici sont celles de descriptions de toutes les relations implicatives de tous les paradigmes complets, un ensemble d'informations excessivement redondant, bien loin de l'optimalité descriptive qui devrait être recherchée lorsque l'on compare des longueurs de description. De plus, aucune information fréquentielle tirée de corpus n'a été pour l'instant utilisée : l'infinitif et la première personne du subjonctif imparfait jouent pour l'instant exactement le même rôle. Cela dit, les fréquences lexicales sont quant à elles directement prises en compte, et on en revient à la problématique, déjà abordée à la section 4.2, du choix de la source d'informations fréquentielles : est-il plus légitime de prendre en compte les fréquences lexicales, les fréquences en corpus, ou une combinaison des deux, et laquelle ?

Une autre question qui méritera d'être approfondie est celle des paradigmes non canoniques, dont certains sont pour l'instant tout simplement ignorés. La surabondance pourrait être en partie gérée en laissant sous-spécifiées certaines composantes de la description qui permettent en principe de spécifier complètement les patrons d'alternance.



(a) Vue générale de la hiérarchie : les macro-classes sont indiquées, mais pas leur contenu n'est pas détaillé.



(b) Gros plan sur la macro-classe correspondant aux verbes en *-ar* (macro-classe grisée dans la sous-figure (a)). Les feuilles correspondent aux micro-classes qui composent cette macro-classe.

FIGURE 4.5 – Classes flexionnelles verbales obtenues pour portugais avec la stratégie de segmentation locale. Les nœuds sont associés au nombre de lemmes qu'ils contiennent.

Mais il se pourrait qu'au moins dans certains cas il faille plutôt considérer, comme en Alexina[PARSLI], que l'on est en présence de (sous-)paradigmes qui doivent être appréhendés séparément. Les paradigmes défectifs ou déficients devront également être modélisés d'une façon adaptée, tout comme les paradigmes hétéroclites. Ce dernier cas est particulièrement intéressant, puisque l'on pourrait songer à construire une hiérarchie de classes flexionnelles qui laisse de côté les comportements flexionnels rarement attestés dans le lexique, puis cherche à rattacher ces derniers aux classes flexionnelles ainsi obtenues, quitte à s'autoriser à combiner plusieurs sous-paradigmes relevant de plusieurs de ces classes.

Sur un plan plus technique, on pourrait penser à utiliser d'autres stratégies de *clustering* (descendant plutôt qu'ascendant, avec une optimisation plus globale que par une séquence de fusions localement optimales, etc.).

Enfin, il reste à tester cette approche sur d'autres langues dont les systèmes morphologiques ont des propriétés et des caractéristiques aussi typologiquement diverses que possible. On pourrait alors comparer les résultats avec les analyses traditionnelles (notamment pour les langues qui ont une tradition grammaticale) et avec les analyses proposées par les linguistes formels et de terrain (y compris pour des langues peu décrites).

Il n'en reste pas moins que les pistes ouvertes par ce travail sont prometteuses, non seulement parce que les hiérarchies de classes obtenues sont proches des classifications traditionnelles, mais également parce qu'elles montrent l'intérêt qu'il y a à combiner une modélisation fondamentalement abstractive et des idées issues d'approches constructives. Nous avions déjà défendu au chapitre 2 l'idée selon laquelle approches abstractives et approches constructives ne sont pas à opposer, mais procèdent de points de vue différents sur les systèmes morphologiques. Faire interagir ces deux points de vue reste un champ de recherche ouvert, mais le travail présenté dans cette section va dans ce sens.

4.4 Bilan et perspectives

Nous avions relevé à la section A.1.3 que la différence fondamentale entre approches abstractives et approches constructives résidait avant tout dans le statut épistémologique accordé à la notion de classe flexionnelle, c'est-à-dire sur la nature de ce que recouvrent les propriétés systémiques du niveau flexionnel : une approche constructive suppose l'existence *a priori* d'un système flexionnel, là où une approche abstractive fait émerger *a posteriori* des propriétés systémiques. En un sens, les trois dernières sections de ce chapitre ont illustré ce que pouvait apporter une approche constructive, puis une approche abstractive, et enfin une approche mixte bien qu'orientée-abstractif. Ceci montre ce que nous écrivions déjà à la section A.1.3 : faire émerger des propriétés systémiques de façon

abstractive n'est pas incompatible avec le fait de conférer à ces propriétés systémiques un statut autonome. On peut penser ici à la création de formes nouvelles, par exemple dans un contexte d'apprentissage de la langue, d'intégration d'emprunts (Walther et Sagot, 2011b) ou de construction de néologismes, ou, en diachronie, pour le remplacement de formes par d'autres, notamment dans le cadre de processus de réfection, c'est-à-dire de remplacement de formes ou ensemble de formes par de nouvelles formes ou ensembles de formes.

Ces différents phénomènes constituent autant de directions de recherche, et peuvent être étudiés sous divers angles : par des approches quantiatives, comme esquissé dans ce chapitre, par des approches psycholinguistiques voire neurolinguistiques, mais également par l'étude de l'évolution diachronique des systèmes linguistiques. Arrêtons-nous quelques instants sur ce dernier angle. C'est un phénomène général dans l'évolution des langues que les paradigmes flexionnels sont souvent refaits, c'est-à-dire que certaines irrégularités, qui résultent souvent de lois phonétiques régulières, sont progressivement effacées par le remplacement de certaines formes par d'autres aboutissant à des paradigmes perçus comme plus réguliers : c'est le phénomène de la réfection analogique, que l'on peut modéliser comme le remplacement de patrons d'alternances rares par des patrons d'alternance plus fréquents et/ou moins complexes [29, 30]. Ce phénomène peut être intraparadigmatique, pour simplifier un paradigme perçu comme trop complexe [31], ou interparadigmatique, pour harmoniser un paradigme irrégulier avec un ensemble numériquement plus important de paradigmes perçus comme réguliers [32]. Ainsi, la réfection analogique fait baisser la complexité du système morphologique, à la fois en un sens constructif (comme à la section 4.1) et en un sens abstractif (comme à la section 4.2). Notons qu'il a été observé depuis longtemps que les formes les plus fréquentes étaient moins susceptibles d'être refaites par analogie : la fréquence protège de la réfection. C'est

29. L'exemple le plus célèbre, dans l'histoire de la linguistique, est peut-être celui du latin archaïque *honos* (NOM)/**honosem* (GEN) HONNEUR, passé par rhotacisme invervocalique régulier du *-s-* à *honos*/*honorem*, puis refait par analogie en *honor*/*honorem* (cf. notamment de Saussure, 1916).

30. Les mécanismes d'analogie ne se limitent naturellement pas à la réfection analogique à l'intérieur des paradigmes flexionnels. Ils sont également au cœur de nombreux processus de création lexicale et notamment du processus de *dérivation inverse* ou *rétroformation* : ce processus consiste à partir d'un lexème existant, à l'identifier à un lexème dérivé d'une base non attestée, et à créer ce lexème base par analogie avec des couples existants constitués d'un lexème de base et d'un lexème qui en est dérivé. On pourra par exemple se rapporter à Garnier (2016) pour une étude approfondie de ce phénomène en latin.

31. C'est le cas du latin *honor* mentionné dans une note précédente. Un autre exemple d'évolution diachronique impliquant plusieurs réfections analogiques intraparadigmatiques est fourni par le grec ancien μήν 'mois'. Comme l'explique en effet Beekes (2009, p. 945), il faut partir d'une forme proto-indo-européenne **meh$_1$n-es-s* 'lune, mois'. À partir d'une forme oblique comme le génitif **mēnsos* (de **meh$_1$nsos*) d'où **mēnnos*, on a refait un nominatif **mēns* par analogie, qui évolue de façon régulière en **mens* (loi d'Osthoff), d'où les variantes dialectales attestées μείς et μής. Un nouveau nominatif μήν a été forgé plus tard sur la base du génitif, passé entre temps à μήνος par simplification régulière de la géminée *-nn-* de l'ancien **mēnnos*.

32. Le même exemple que dans la note précédente peut être ici repris : à partir du génitif μήνος, et par analogie avec le nominatif Ζεύς 'Zeus' et son génitif Ζηνός, le dialecte grec éléen a formé un nouveau nominatif μεύς (Beekes, 2009, 945).

là du reste un constat qui pourrait être de nature à guider la façon dont on veut représenter les systèmes morphologiques et mesurer leur complexité.

La dichotomie entre approches constructives et abstractives peut alors s'instancier comme suit : ces phénomènes de réfection analogiques reposent-ils sur un système de classes flexionnelles disponible *a priori* et dans lequel ils cherchent à faire rentrer des paradigmes nouveaux ou trop irréguliers ? Ou bien les phénomènes de réfection analogique, qui se passeraient alors avant tout de forme à forme, ont-ils pour conséquence l'émergence ou la généralisation de régularités que l'on peut interpréter *a posteriori* au moyen de cette abstraction qu'est la notion de classe flexionnelle ?

La réponse à cette question est délicate à plusieurs égards. Tout d'abord, il n'est pas évident d'identifier des tests qui permettraient de trancher entre ces deux hypothèses. Ensuite, ces points de vue pourraient bien capturer tous deux une partie des phénomènes. Il n'y a rien d'incompatible à ce que d'une part certaines formes soient refaites par analogie sur la base de formes individuelles [33], et d'autre part certains paradigmes soient déformés pour les faire rentrer dans une classe flexionnelle dont l'existence en soi serait disponible pour ce faire. Mais surtout, la différence entre ces deux points de vue n'est pas aussi nette qu'on pourrait le penser à première vue. L'étude des patrons d'alternance impliquant plus de deux cases pourrait par exemple fournir des indices permettant le développement de modèles intermédiaires (Bonami et Beniamine, 2015). Il en est de même du développement de modèles de segmentation semi-locaux, tels que définis plus haut, couplés à une structuration fine des paradigmes, comme c'est le cas en un certain sens dans Alexina[PARSLI]. L'étude quantitative de l'évolution diachronique des systèmes flexionnels, non pas au niveau des formes individuelles mais à celui des systèmes dans leur ensemble, pourrait apporter des éléments importants dans cette réflexion.

33. L'exemple de la note précédente est ici intéressant : Ζεύς (NOM)/Ζηνός (GEN) 'Zeus' est le modèle ayant servi au dialecte grec éléen à former un nouveau nominatif μεύς 'mois' à partir du génitif μήνος, bien que Ζεύς soit le seul lemme à se fléchir de la sorte. Pour prendre un autre exemple relevant de la morphologie dérivationnelle et non flexionnelle peut citer le grec ancien βασίλισσα 'reine', l'une des formes féminines de βασιλεύς 'roi' (du grec mycénien *qa-si-re-u* /g^{w}asileus/), dont l'étymologie est incertaine, tant concernant le radical que le suffixe (Buck, 1914 ; Beekes, 2009). Pourtant, ce suffixe -ισσα, perçu à tord ou à raison comme un suffixe de féminin et qui n'était présent au départ dans cette seule forme, a été étendu par analogie à quelques autres formes en grec même, emprunté en latin, hérité en français puis emprunté au français par l'anglais, chaque étape le faisant gagner en productivité, d'où des mots anglais comme ACTRESS 'actrice', LIONESS 'lionne', PRIESTESS 'prêtresse', etc.

Chapitre 5

Lexiques syntaxiques : modélisation, implémentation et développement [1]

Sommaire

Les informations lexicales syntaxiques, au premier rang desquelles les informations de valence, ont fait l'objet de travaux depuis plusieurs décennies, comme décrit aux sections A.5 et A.6, avec deux objectifs principaux : la description linguistique et le développement d'analyseurs syntaxiques symboliques. En effet, dans ce type d'analyseurs, disposer d'informations syntaxiques est une nécessité, soit parce qu'ils reposent sur des formalismes qui représentent la quasi-totalité des informations

1. Les travaux décrits dans ce chapitre ont été réalisés pour une part en collaboration avec Laurence Danlos (Université Paris 7).

linguistiques dans le lexique (ainsi les grammaires d'adjonction d'arbres [2]), soit parce qu'ils répartissent de façon plus équilibrée l'information linguistique entre une grammaire et un lexique (ainsi les grammaires lexicales-fonctionnelles [3]). Comme nous le verrons plus en détails au chapitre 9, l'avènement des analyseurs syntaxiques statistiques et désormais neuronaux ne remet pas en cause la pertinence de l'information stockée dans les lexiques syntaxiques, information complémentaire de celle fournie par les corpus arborés. Certes, le couplage de tels analyseurs avec des lexiques syntaxiques est moins direct et plus difficile. C'est l'une des raisons pour lesquelles de tels couplages ont fait l'objet de peu de travaux, l'autre étant naturellement que, même sans exploiter de lexique syntaxiques, de tels analyseurs ont progressivement atteint des performances très satisfaisantes. Nous reviendrons sur ces questions au chapitre 9, où nous montrerons notamment que la prise en compte d'informations lexico-syntaxiques dans des analyseurs probabilistes, par hybridation symbolique/statistique ou par des mécanismes de réanalyse permet d'améliorer les performances.

Le développement de lexiques syntaxiques, indispensable il y a quelques années, restent ainsi des enjeux importants. C'est d'autant plus le cas avec l'intérêt renouvellé, depuis quelques années, des analyseurs syntaxiques dits profonds, qui, comme le faisaient nombre d'analyseurs symboliques depuis longtemps, sont capables de fournir des analyses prenant en compte des phénomènes tels que le contrôle et la montée, phénomènes éminemment lexico-syntaxiques.

Mais développer des lexiques syntaxiques suppose de disposer d'un modèle de représentation des informations lexico-syntaxiques. La pertinence linguistique de tels modèles est une condition nécessaire à leur applicabilité à des langues diverses, à l'utilisation des lexiques associés à des fins de linguistique descriptive, théorique et quantitative, mais également à leur intégration à des analyseurs syntaxiques, tant symboliques — et ce, en étant aussi indépendant que possible du formalisme choisi — que statistiques ou neuronaux, et *a fortiori* hybrides.

Dans ce chapitre, nous nous pencherons ainsi successivement sur la modélisation de l'information lexico-syntaxique, en motivant et présentant le niveau syntaxique de notre formalisme lexical Alexina, puis sur le développement de lexiques syntaxiques, en présentant nos travaux sur le développement du niveau syntaxique du lexique Alexina du français, le Le*fff*. Plus précisément, nous présenterons tout d'abord brièvement deux des principales ressources lexico-syntaxiques pour le français développées depuis plusieurs décennies, les tables du Lexique-Grammaire et Dicovalence, dont nous nous sommes insiprés dans nos travaux mais dont nous montrerons les limites. Nous présenterons alors notre propre modèle lexico-syntaxique, qui constitue le niveau syntaxique de notre modèle lexical Alexina, dont le niveau morphologique a fait l'objet des chapitres 2 à 4.

2. Tree Adjoining Grammars (TAG ; Joshi *et al.*, 1975 ; Joshi, 1987).
3. Lexical-Functional Grammars (LFG ; Bresnan, 1982 ; Kaplan et Bresnan, 1982)

Nous discuterons à la fois des notions sous-jacentes et de la façon dont nous représentons les entrées simples [4] et les entrées complexes [5], en les illustrant au moyen du Le*fff*, notre lexique à grande échelle du français, puis en discutant de l'applicabilité du modèle à d'autres langues. Nous décrirons ensuite les différentes approches que nous avons mises en œuvre pour développer le Le*fff*, lexique syntaxique riche et à large couverture. Nous discuterons brièvement des moyens dont on peut évaluer une telle ressource, avant de finir sur quelques réflexions générales et pistes pour des travaux futurs.

5.1 Modélisation de l'information lexico-syntaxique

5.1.1 Le cas du français : Lexique-Grammaire et Dicovalence

Comme décrit plus amplement à la section A.6, les deux principales ressources lexicales syntaxiques pour le français sont les tables du Lexique-Grammaire (Gross, 1975 ; Boons *et al.*, 1976a,b ; Guillet et Leclère, 1992) et le dictionnaire Dicovalence (van den Eynde et Mertens, 2006), toutes deux développées depuis de nombreuses années dans des laboratoires de linguistique. Une troisième ressource d'importance, elle aussi mentionnée précédemment, est le dictionnaire syntaxico-sémantique de Dubois et Dubois-Charlier (1997), *Les Verbes Français* (LVF). Le choix des propriétés syntaxiques renseignées pour chaque entrée et la façon, relativement obscure, dont elles sont codées dans LVF ne reposent pas véritablement sur un modèle spécifique de la syntaxe lexicale. En réalité, LVF est une ressource inspirée par le Lexique-Grammaire, dont il simplifie les principes tout en les prolongeant au niveau sémantique. Pour cette raison, nous laissons pour l'instant LVF de côté, mais nous y reviendrons au chapitre suivant, consacré au développement du Le*fff*, en tant qu'il a été l'une des ressources utilisées à cette fin.

Deux stratégies différentes ont été mises en œuvre dans le développement de ces ressources. Dicovalence se concentre volontairement sur les verbes les plus fréquents (3738 lemmes), et, pour ces verbes, sur leurs emplois les plus fréquents (8313 entrées). *A contrario*, le développement du Lexique-Grammaire a toujours été orienté vers un objectif pourtant illusoire d'exhaustivité, aboutissant à 6 500 lemmes décrits par 13 375 entrées [6].

4. Les « lexèmes » au sens de la Théorie Sens-Texte (Meľčuk, 1974, 1988 ; cf. A.5).

5. Les « locutions » au sens de la Théorie Sens-Texte.

6. À titre d'illustration, dans le Lexique-Grammaire, la Table 31H N_0^{hum} *V* (verbes intransitifs à sujet humain possible mais sujet phrastique impossible) comportait 129 entrées dans l'annexe de Boons *et al.* (1976a) ; cette même table comporte aujourd'hui 626 verbes, dont certains lemmes peu usités tels que *bovaryser*, *calancher*, *se curedenter*, ou *faonner*. Au niveau des entrées, la même exhaustivité est recherchée : dans la Table 9 N_0 *V* (*Que P*)$_1$ *à* N_2 (verbes ditransitifs à sujet phrastique et objet direct phrastiques possibles et à objet indirect introduit par *à*), inclut des verbes tels que *bourdonner*, *bramer*, *chuinter*, *coasser*, *couiner*, *croasser*, *crépiter*, ou *gargouiller* (ces entrées sont analysées par « fusion » : *bramer* = *dire en bramant*, d'où *bramer quelque chose à quelqu'un*). D'une manière plus générale, les verbes peu usités et les entrées douteuses (correspondant à des phrases traditionnellement préfixées par « ? » dans la littérature linguistique) sont considérés comme acceptables par le Lexique-Grammaire, alors que Dicovalence aura tendance à ne pas les prendre en compte.

D'un point de vue méthodologique, le Lexique-Grammaire repose sur une structuration hiérarchique reflétée par une organisation en tables. À l'inverse, Dicovalence est un ensemble non structuré d'entrées. La structuration des entrées, que l'on peut formaliser par un graphe d'héritage, n'a aucune conséquence pour la réalisation d'analyseurs syntaxiques. Néanmoins, une telle structuration reflète des généralisations linguistiques pertinentes et peut faciliter le développement et la maintenance de ressources lexicales.

Le Lexique-Grammaire et Dicovalence sont deux ressources très riches mais incomplètes. Toutefois, elles peuvent mutuellement s'enrichir. À titre d'exemple, Dicovalence comporte des informations précises sur ce qui y est appelé « pronoms suspensifs » (pronoms interrogatifs et adverbes interrogatifs ou indéfinis, comme *à qui* ou *quand*), tandis que le Lexique-Grammaire ne prend pas en compte les interrogatives indirectes [7] et regroupent sous l'identifiant ADV les compléments pronominalisables en *ainsi, autant, (Prép) quand*, etc. À l'inverse, le Lexique-Grammaire comporte des informations plus précises que Dicovalence sur certains points, comme le système des noms parties du corps [8], les compléments sous-catégorisés mais non pronominalisables, certaines *restructurations* [9]

Enfin, certaines propriétés et constructions complexes font l'objet de codages différents dans chacune des ressources. On pourra par exemple se reporter à (Sagot et Danlos, 2009) pour une étude comparative sur les constructions pronominales, qui sert de préalable à leur traitement dans le L*efff* (cf. par exemple Sagot et Danlos, 2009).

5.1.2 Motivations pour le développement d'un nouveau formalisme syntaxique

Aucune de ces deux approches, et par conséquent aucune de ces deux ressources lexicales, ne saurait donner satisfaction :

- le Lexique-Grammaire ne propose pas de typage des arguments syntaxiques (contrairement à Dicovalence) ;
- Dicovalence ne traite qu'imparfaitement certains types de réalisations syntagmatiques des arguments, et notamment les réalisations phrastiques ;
- certaines informations cruciales (contrôle, montée, attribution) sont seulement partiellement codées ;
- les informations syntaxiques sont associées à des entrées, sans prendre en compte les spécificités de certaines formes fléchies à cet égard (sujet devenant facultatif pour une infinitive, par exemple) ;

7. Voir cependant Nakamura (2006), qui a codé les interrogatives indirectes pour la Table 6 ($N_0\ V\ (Que\ P)_1$).

8. Il permet de rendre compte d'alternances telles que *Luc caresse les cheveux de Marie / Luc lui caresse les cheveux*.

9. Un exemple : *Luc copie les habitudes de Léa / Luc copie Léa dans ses habitudes*, ou encore certaines relations de dérivation morphologique.

— Lexique-Grammaire n'est pas utilisable directement dans des applications automatiques ;

— Dicovalence ne couvre qu'une sous-partie du seul lexique verbal, et le Lexique-Grammaire ne contient pas d'entrées pour certains des verbes les plus fréquents (semi-auxiliaires, par exemple), tout en considérant comme possible de nombreuses constructions dont l'acceptabilité est très douteuse.

Nous avons donc développé un modèle lexico-syntaxique original, qui constitue le niveau syntaxique de notre formalisme lexical Alexina, dont le niveau morphologique a fait l'objet du chapitre 2. Ce faisant, nous nous sommes inspirés des travaux antérieurs sur la modélisation de l'information lexico-syntaxique, dont nous proposons une discussion critique à la section A.5. En particulier, nous avons essayé de reprendre à notre compte certains aspects des modèles utilisés par le Lexique-Grammaire et par Dicovalence. Une autre source d'inspiration a été le formalisme LFG (cf. section A.5), dont on reconnaîtra, quoique transformées, certaines conventions d'écriture. L'objectif du niveau syntaxique d'Alexina est multiple :

— permettre de représenter les informations syntaxiques pertinentes à la fois pour des usages en linguistique et pour des applications de traitement automatique comme l'analyse syntaxique, et notamment l'analyse syntaxique symbolique ;

— représenter ces informations de façon aussi indépendante que possible de tel ou tel formalisme particulier, y compris LFG, tout en préservant la pertinence linguistique du modèle [10] ;

— pouvoir modéliser autant que possible toute la diversité des types d'entrées et des types d'informations possibles, sans laisser de côté certains phénomènes spécifiques, surtout s'ils sont fréquents (entrées pour les verbes modaux, réalisations phrastiques…) ;

— distinguer le niveau de l'entrée lexicale et celui de ses différentes formes fléchies, afin de pouvoir représenter les spécificités syntaxiques de ces dernières ;

— disposer d'un modèle permettant de traiter des constructions moins usuelles, telles que les constructions impersonnelles ou moyennes.

10. Prenant comme exemple les analyseurs syntaxiques, certains lexiques Alexina, et plus particulièrement le Le*fff*, le lexique Alexina du français, sont intégrés d'une façon ou d'une autre à des analyseurs syntaxiques reposant sur des formalismes aussi divers que les TAG lexicalisées (produites à partir de descriptions plus abstraites appelées méta-grammaires ; analyseur FRMG) (Thomasset et Villemonte de La Clergerie, 2005 ; Villemonte de La Clergerie *et al.*, 2009a ; Villemonte de La Clergerie, 2013), les grammaires lexicales-fonctionnelles (LFG ; analyseur SxLfg, Boullier et Sagot, 2005 ; Boullier *et al.*, 2005b ; Sagot, 2006 ; Sagot et Boullier, 2006), les grammaires d'interaction (Guillaume et Perrier, 2010a), les grammaires de pré-groupes (Béchet et Foret, 2009), mais également les grammaires non contextuelles probabilistes à annotations latentes (Candito et Seddah, 2010 ; Sigogne et Constant, 2012) ou les grammaires de dépendance probabilistes (Candito *et al.*, 2010 ; Urieli et Tanguy, 2013). Un système hybride particulièrement performant, couplant FRMG et analyseur statistique en dépendances, repose également sur le Le*fff* (Villemonte de La Clergerie, 2014).

5.1.3 Le formalisme lexical Alexina : le niveau syntaxique

C'est dans cet esprit que nous avons développé le niveau syntaxique d'Alexina, que nous allons présenter dans la suite de ce chapitre. Nous allons procéder en deux étapes. Tout d'abord, nous justifierons et définierons les concepts sous-jacents à partir des entrées simples, dont nous commenterons quelques exemples issus du Le*fff*. Puis nous montrerons quelles classes d'entrées complexes peuvent être représentées dans Alexina dans sa version actuelle, là aussi en illustrant notre propos sur des exemples tirés du Le*fff*. En effet, le Le*ffff* est à ce jour la seule ressource Alexina disposant d'informations syntaxiques de qualité et à grande échelle (Sagot *et al.*, 2006 ; Sagot, 2010, 2013b). Comme indiqué précédement, le lexique Alexina de l'espagnol, le Le*ff*e, dispose également d'informations syntaxiques, mais de façon très imparfaite et préliminaire. Pour cette raison, et bien que le niveau syntaxique d'Alexina soit conçu, comme son niveau morphologique, pour être adapté à la description du lexique de langues variées, la suite de ce chapitre se concentre sur le français et sur le lexique Le*fff* [11].

5.1.3.1 Fonctions syntaxiques, réalisations et redistributions

Un lexique Alexina est constitué d'entrées lexicales qui, dans le cas des entrées simples, correspondent chacune à un lexème. Le cas des entrées complexes et des locutions sera abordé plus bas. Un lexème correspond à un sens particulier d'un lemme ayant des propriétés morphologiques et syntaxiques cohérentes. Il y a donc par exemple dans le Le*fff* deux entrées distinctes pour le lemme *voler*. En ce sens, un lexique Alexina combine donc les niveaux morphologique, syntaxique et sémantique.

Au niveau morphologique, une entrée est donc composée d'un lemme, défini par la donnée d'une forme de citation et d'une classe flexionnelle, ou d'un schème au sens d'Alexina$_{\text{PARSLI}}$ (cf. chapitre 2), ce qui lui confère une catégorie morphologique.

Au niveau syntaxique proprement dit, chaque entrée doit décrire quels sont les arguments syntaxiques possibles ou obligatoires du lexème et quelles sont les propriétés de ces arguments et de leurs réalisations possibles. Il est donc nécessaire de définir ce qu'est un argument syntaxique. Bien que le formalisme Alexina puisse s'accommoder de définitions différentes, nous définissons un *argument syntaxique* comme étant un syntagme ou un pronom dont l'existence, le type, les propriétés morphosyntaxiques et la distribution sont contrôlés par une forme donnée. Dans des langues telles que le français ou l'anglais, une forme susceptible d'avoir des arguments, ou *forme prédicative*, peut notamment être une forme verbale, nominale ou adjectivale. Dans la plupart des cas, les

11. Les similarités entre le niveau syntaxique d'Alexina et la norme ISO LMF (cf. plus bas), mais également certains points communs avec le formalisme LFG, sont plutôt rassurantes quant à son adéquation à la description de langues variées. Outre l'espagnol, des travaux très préliminaires ont été réalisés sur le persan, sans que le modèle ne semble rencontrer d'obstacle majeur.

arguments syntaxiques d'une forme correspondent aux arguments sémantiques, c'est-à-dire aux actants du procès dénoté par la forme [12]. Toutefois, un argument syntaxique peut ne pas correspondre à un actant : c'est alors un pseudo-argument. C'est par exemple le cas en français du *se* (ou *s'*) de verbes pronominaux tels que *s'évanouir*. À l'inverse, il est parfois impossible de réaliser en syntaxe un des actants. C'est par exemple le cas de l'agent dans les constructions dites *se*-moyen (ou à agent fantôme : *les crevettes se mangent avec les doigts *(par les bretons)* ; Abeillé, 2002, p. 193).

L'ensemble des arguments syntaxiques d'une forme donnée et les contraintes qui sont associées à ces arguments sont modélisés au moyen d'un *cadre de sous-catégorisation*. Un tel cadre code donc l'ensemble des arguments syntaxiques, complété autant que possible par des propriétés syntaxiques complémentaires qui sont souvent dépendantes de la langue, telles que des informations sur les phénomènes de contrôle ou d'attribution, sur le mode des complétives, le gouverneur syntaxique requis (par exemple pour les noms prédicatifs faisant usage de verbes supports), etc. Ainsi, pour un lexème comme *permettre*, il est loisible de spécifier qu'une éventuelle complétive objet doit être au subjonctif, et que le sujet d'une éventuelle infinitive objet correspond à l'objet indirect, s'il est exprimé : *La ponctualité du train permet que les passagers attrapent leurs correspondances sans difficulté*, ou *Le retard du train permet à Marie de partir malgré tout*.

Dans un cadre de sous-catégorisation, chaque argument est modélisé comme suit : d'une part, il est associé à une *fonction syntaxique*, c'est-à-dire un ensemble cohérent de contraintes morphologiques et syntaxiques, dont les plus habituelles sont les fonctions sujet et objet direct ; d'autre part, il indique l'ensemble de ses *réalisations* possibles, pronominales ou syntagmatiques. Les pseudo-arguments sont également inclus dans le cadre de sous-catégorisation, où l'on indique leur réalisation, mais ne reçoivent pas de fonction syntaxique.

La notion de fonction syntaxique est courante dans de nombreux formalismes et approches (Tesnière, 1959 ; Kaplan et Bresnan, 1982 ; Perlmutter et Postal, 1983). Pour notre part, nous définissons les fonctions syntaxiques de façon spécifique à chaque langue au moyen de critères dont on peut évoquer brièvement les principaux comme suit :

— le principe de commutation, en prenant en compte les pronoms comme les syntagmes, contrairement aux choix fait par les auteurs du lexique de valence verbale Dicovalence (van den Eynde et Mertens, 2003, 2006), qui se limitent aux seuls pronoms : si un pronom ou un syntagme peut être remplacé à la même position par un autre pronom ou syntagme de façon mutuellement exclusive, et ce sans changer la structure de dépendance sous-jacente, alors ils occupent la même fonction syntaxique ;

12. Alexina se limitant à la morphologie et à la syntaxe, nous laissons de côté la question de l'obligatorité sémantique des arguments à réalisation non obligatoire en syntaxe. En revanche, un argument sémantique, même obligatoire, peut ne pas être obligatoirement réalisé en syntaxe.

— le principe de réalisation unique : pour une occurrence donnée d'une forme prédicative, chacun de ses arguments ne peut être réalisé qu'au plus une fois.

Avec ces critères, la correspondance entre fonctions sémantiques et syntaxiques n'est pas nécessairement unique. Plusieurs cadres de sous-catégorisation distincts peuvent devoir être associés à différentes formes d'un même lexème, pour au moins deux raisons : d'abord, certaines formes ont des comportement spécifiques au niveau syntaxique (par exemple, en français, l'infinitif d'un verbe peut ne pas avoir de sujet réalisé). Ensuite, les formes fléchies d'un lexème donné peuvent mettre en correspondance arguments syntaxiques et sémantiques de différentes façons, y compris pour une même forme : c'est le phénomène bien connu des alternances syntaxiques régulières, dont les constructions passives et impersonnelles constituent deux exemples. Parmi elles, nous posons qu'il existe une correspondance non marquée entre actants sémantiques et fonctions syntaxiques qui correspond à des constructions syntaxiques non marquées ; cette correspondance est disponible pour la majorité des lexèmes, mais pas nécessairement pour tous. Pour le français, par exemple, il s'agit de ce que l'on dénote généralement par la notion de voie (ou diathèse) active.

À la suite de travaux antérieurs (Perlmutter et Postal, 1983 ; Candito, 1999), nous opposons alors la notion de fonction syntaxique décrite jusqu'ici, que nous qualifierons parfois de *fonction syntaxique finale*, à une notion de *fonction syntaxique initiale*, qui en abstrait l'effet des alternances syntaxiques. Plus précisément, nous définissions la fonction syntaxique initiale d'un argument syntaxique comme étant la fonction syntaxique finale de ce même argument dans un contexte non marqué. [13]. En conséquence, l'ensemble des fonctions syntaxiques initiales possibles est inclus dans celui des fonctions syntaxiques finales. Ceci permet d'associer à un lexème donné un cadre de sous-catégorisation initial unique, qui contient un ensemble de fonctions syntaxiques initiales associées aux réalisations possibles et aux propriétés ou contraintes syntaxiques associées qui sont valides dans les situations non-marquées.

Les correspondances marquées (par exemple le passif en français) ou les correspondances non marquées utilisées dans des contextes syntaxiques marqués (par exemple les constructions impersonnelles du français) sont alors définies comme des *redistributions* de l'ensemble des fonctions syntaxiques initiales vers un ensemble de fonctions syntaxiques finales qui en diffèrent en tout ou partie, et qui sont éventuellement associées à des propriétés syntaxiques spécifiques. Chaque redistribution doit être définie pour une langue donnée de façon formelle, en spécifiant la façon dont elle attribue à certains argu-

13. Un autre modèle auquel le nôtre pourrait se comparer est celui proposé par Bonami (1999), qui combine des informations de valence liées à chaque lexème avec des informations provenant de *schémas argumentaux* qui ont vocation à les compléter ou à les amender. L'objectif est toutefois différent, puisque Bonami (1999) cherche avant tout à rendre compte des propriétés syntaxiques et sémantiques d'arguments prépositionnels qui sont facultatifs à la fois en syntaxe et en sémantique.

ments syntaxiques des fonctions syntaxiques finales distinctes de leurs fonctions syntaxiques initiales, les changements qu'elle opère dans l'inventaire de réalisations possibles de chaque argument et les propriétés syntaxiques associées (contrôle, optionalité, etc.), et les propriétés ou contraintes syntaxiques supplémentaires qu'elle induit. Enfin, chaque redistribution peut être amenée à agir de façon différenciée selon les formes d'un même lexème, soit en affectant la construction du cadre de sous-catégorisation final (par exemple, en rendant facultatif le sujet d'un lexème verbal dès lors que la forme considérée est infinitive), soit en bloquant son application (par exemple, la redistribution passive ne peut s'appliquer qu'à une forme de participe passé) ; c'est le rôle des *marqueurs morphosyntaxiques* produits par la grammaire morphologique, en parallèle aux structures de traits morphosyntaxiques, de faire transiter les informations nécessaires du niveau morphologique au niveau syntaxique [14].

Il y a donc une double transformation entre les informations associées à un lexème et les informations associées à l'une de ses formes : une transformation morphologique, la flexion, et une transformation syntaxique, la redistribution. On peut alors qualifier de *lexique intensionnel* un lexique Alexina tel que nous l'avons décrit jusqu'ici. Dans un lexique intensionnel, une entrée correspond donc à un lexème, c'est-à-dire une forme canonique, une classe flexionnelle (ou un schème), un cadre de sous-catégorisation initial et une liste de redistributions possibles. Un tel lexique peut être compilé automatiquement, grâce aux descriptions formelles des redistributions et des classes flexionnelles, en un lexique où chaque entrée correspond à la donnée d'une forme fléchie et de l'application d'une redistribution qui lui est compatible. Un tel lexique est qualifié de *lexique extensionnel*. Naturellement, les travaux de développement ou de correction automatiques ou semi-automatiques de lexiques ciblent le niveau intensionnel. À l'inverse, les outils automatiques tels que les étiqueteurs morphosyntaxiques ou les analyseurs morphologiques ou syntaxiques utilisent directement les lexiques de niveau extensionnel. Cette dichotomie entre lexique intensionnel et extensionnel et donc entre entrées intensionnelles et extensionnelles constitue l'architecture à deux niveaux d'Alexina. Elle se situe conjointement au niveau morphologique (lemme vs. forme fléchie) et au niveau syntaxique (sous-catégorisation initiale vs. finale), permettant la modélisation explicite des interactions entre ces deux niveaux via les marqueurs morphosyntaxiques.

14. On pourra se reporter à (Vernerová *et al.*, 2014) pour un rapide panorama des ressources lexico-syntaxiques qui explicitent les informations de redistribution, quelle que soit la façon de le faire. On constatera, avec les auteurs, qu'elles sont finalement bien moins nombreuses qu'on pourrait s'y attendre. Il nous semble qu'il s'agit là d'une conséquence du fait que les redistributions moins fréquentes ont tendance à être ignorées alors que les redistributions plus fréquentes sont souvent considérées comme automatiquement disponibles dès lors que telle ou telle condition est remplie (par exemple, en considérant tout verbe transitif direct comme passivable et inversement), jugeant ainsi que le lexique n'a pas besoin de contenir de telles informations. Ce point de vue est pourtant clairement erroné, comme rappelé par (Vernerová *et al.*, 2014) et comme illustré par le caractère non-passivable du verbe transitif direct *regarder* (au sens de *concerner*) et par le caractère passivable du verbe transitif indirect *obéir*.

Il est intéressant de noter que l'extension syntaxique de la norme ISO pour l'encodage des informations lexicales, LMF (Lexical Markup Framework) Francopoulo *et al.* (2006), est très similaire au modèle Alexina, qui lui est pourtant antérieur. C'est d'ailleurs grâce à cette correspondance que les tables du Lexique-Grammaire du français, préalablement converties dans le modèle Alexina (Tolone et Sagot, 2009 ; Tolone, 2011), ont pu être ensuite encodées en LMF (Laporte *et al.*, 2013).

5.1.3.2 Entrées simples : illustration sur le Le*fff*

Le lexique Alexina le plus important est le Le*fff*, Lexique des Formes Fléchies du Français, est un lexique morphologique et syntaxique à grande échelle pour le français développé dans le cadre d'Alexina au moyen de diverses techniques semi-automatiques, mais également grâce à un lourd travail manuel reposant et sur l'étude de corpus et sur l'introspection. Le développement du Le*fff* est brièvement décrit ci-dessous à la section 5.2.1. Le Le*fff* contient plus de 110 000 entrées intensionnelles (lexèmes) qui couvrent près de 550 000 entrées de niveau forme fléchie réparties dans toutes les catégories. Nous évoquerons la problématique de l'évaluation d'un tel lexique à la section 5.3.

Les fonctions syntaxiques sont définies dans le Le*fff* par des critères proches de ceux de Dicovalence (van den Eynde et Mertens, 2006). L'inventaire des fonctions syntaxiques est le suivant : *Suj* (sujet), *Obj* (objet direct), *Objà* (objet indirect introduit canoniquement par la préposition *à*), *Objde* (objet indirect introduit canoniquement par la préposition *de*), *Loc* (locatif) [15], *Dloc* (délocatif), *Att* (attribut) [16], *Obl*, *Obl2* ou *Obl3* (autres arguments obliques). Les critères définitoires de ces fonctions sont décrits dans (Sagot et Danlos, 2007).

Dans le Le*fff*, les différentes réalisations qui peuvent instancier une fonction syntaxique sont de trois types :

15. Dans l'état actuel du Le*fff*, la fonction *Loc* regroupe ce que Bonami (1999) appelle les locatifs (strictement) directionnels (par exemple en *vers*), les locatifs de type but dynamique (par exemple en *jusqu'à*) et les locatifs statiques, c'est-à-dire compatibles avec la copule (par exemple *à*, qu'ils soient à sémantique directionnelle ou non). Ainsi, le Le*fff* ne modélise pas le fait que ces trois types d'arguments locatifs ne sont pas nécessairement tous compatibles avec une entrée lexicale admettant un argument locatif (Bonami, 1999). En revanche, en conférant au sens premier du verbe *aller* un argument locatif obligatoire, cela permet de modéliser le fait que pour ce lexème la réalisation en syntaxe d'un et un seul argument locatif relevant de l'un quelconque de ces trois types est obligatoire.

16. Dans l'état actuel du Le*fff*, la fonction *Att* recouvre plusieurs types d'arguments, qu'il pourrait convenir de distinguer. Il s'agit (i) des « vrais » attributs, qu'ils soient du sujet (*Jean est grand*) ou de l'objet *Marie trouve Jean grand*), (ii) des pseudo-objets, comme *5 euros* dans *Pierre a payé ce livre 5 euros* (cf. *Pierre les a payé de bonne grâce*, et des infinitives à sujet coréférent à un argument autre que le sujet (*Pierre regarde Marie partir*), où l'infinitive est représentés comme un attribut de l'objet, en tant que son sujet coréfère à l'objet. On notera toutefois que nombreux sont les verbes où un « vrai » attribut commute avec une telle infinitive, tout en s'excluant mututellement (cf. *Pierre semble partir/absent*, *Pierre estime Marie (être) intelligente*). Autrement dit, la situation n'est pas aussi tranchée que l'on pourrait le penser de prime abord. Un travail spécifique est ici nécessaire.

— pronoms clitiques,

— syntagme direct : syntagme nominal (*sn*), adjectival (*sa*), infinitif (*sinf*), phrastique fini (*scompl*), interrogative indirecte (*qcompl*) [17] ;

— syntagme prépositionnel : syntagme direct précédé d'une préposition, comme *de-sn*, *à-sinf* ou *pour-sa* ; *à-scompl* et *de-scompl* représentent les réalisations en *à/de ce que P*.

Enfin, comme pour tout lexique Alexina, une fonction dont la réalisation est facultative voit sa liste de réalisations possibles mise entre parenthèses.

Des informations syntaxiques complémentaires (contrôle, mode des complétives, etc.) sont notées dans Alexina de l'une ou l'autre des deux façons suivantes :

— par un couple attribut-valeur : par exemple, *cat=v* indique une catégorie lexicale verbale (qui peut être distincte de la catégorie syntaxique et/ou de la catégorie morphologique [18] ;

— par une *macro* spécifique à chaque lexique : le Le*ffff* contient ainsi par exemple des macros telles que *CtrlSujObj* (contrôle du sujet sur l'objet) ou *ComplSubj* (une réalisation complétive de l'objet direct est nécessairement au subjonctif) ; l'interprétation formalisée de ces macros dépend du contexte d'utilisation, mais une modélisation de ces macros en LFG est fournie avec le Le*ffff*.

Ceci étant posé, nous pouvons illustrer le modèle lexical Alexina et la façon dont il est utilisé dans le Le*ffff* sur un exemple. Soit le lemme verbal *clarifier*. Ce lemme correspond à deux entrées distinctes, l'une pour le sens 'rendre (une idée) plus compréhensible' et l'autre pour le sens 'rendre (un fluide) transparent, clair, pur'. Considérons le premier de ces deux lexèmes, identifié par la forme de citation *clarifier* et un indice sémantique qui le distingue du second, soit CLARIFIER$_1$. Il s'agit d'un lexème transitif (son cadre de sous-catégorisation initial comporte un sujet et un objet direct, qui tous deux peuvent être réalisés de différentes façons). Outre la redistribution active, CLARIFIER$_1$ admet des réalisations marquées, comme le passif ou le *se*-moyen. L'entrée intensionnelle (simplifiée) correspondante, que nous glosons ci-dessous, est la suivante :

17. À ce jour, les réalisations phrastiques introduites par *si* sont réputées couvertes par la réalisation *qcompl*. La question reste toutefois ouverte de savoir s'il conviendrait d'en faire un type spécifique de réalisation.

18. Par exemple, le Le*ffff* comporte des entrées pour des unités à comportement clitique affixal, qui sont très difficiles à classer et induisent, sauf à les modéliser explicitement, à une grande variété de mots inconnus. C'est par exemple le cas du préfixe *ex-* : *ex-président (de la société)*, *ex-petite amie*, *ex-Yougoslavie*. Au niveau morphologique, *ex-* est naturellement invariable, il n'est donc pas porteur d'une catégorie morphologique. Sa catégorie lexicale est quant à elle une catégorie adjectivale, puisqu'il commute avec l'adjectif *ancien*. Enfin, sa catégorie syntaxique est spécifique à son fonctionnement de préfixe (par exemple, on ne peut que l'antéposer, et pas le modifier) : il s'agit de la catégorie *adjPref*, dont une grammaire pourra spécifier le comportement particulier et différent de celui de la catégorie des adjectifs standard.

CLARIFIER$_1$ *v-er :std* Lemma ;v ;<*Suj* :*cln* |*scompl* |*sinf* |*sn*,*Obj* :(*cla* |*scompl* |*sn*)> ;
cat=v, *ComplInd* ;
%actif,*%passif*,*%se_moyen*,*%se_moyen_impersonnel*,
%passif_impersonnel,*%ppp_employé_comme_adj*

Cette entrée décrit un verbe transitif qui se conjugue comme un verbe standard du premier groupe (*v-er :std*) et dont les arguments ont les fonctions syntaxiques Suj (sujet, réalisable comme un clitique nominatif, une complétive, une infinitive ou un syntagme nominal) et Obj (objet direct, réalisable comme un clitique accusatif, une complétive à l'indicatif ou un syntagme nominal) . Les redistributions autorisées sont ici les redistributions active, passive, *se*-moyenne, *se*-moyenne impersonnelle (*il s'est clarifié que…*), passive impersonnelle (*la chose s'est clarifiée lorsque…*) et participe passé employé comme adjectif. Cette entrée va donner naissance à de nombreuses entrées de niveau extensionnel, comme l'entrée ci-dessous, qui correspond à la forme fléchie du participe passé masculin pluriel (étiquette *@Kmp*) pour la redistribution passive. Cette redistribution, formalisée explicitement à part, indique en effet qu'elle s'applique uniquement aux formes dont le marqueur morphosyntaxique est *PastParticiple*, marqueur associé par la grammaire morphologique à toutes les formes de participe passé.

clarifiés v [pred='clarifier$_1$ <*Suj* :*cln*|*scompl*|*sn*,*Obl2* :(*par-sn*)>',
cat=v,*passif*,*pers*,*@Kmp* ;
PastParticiple
%passif

Il en va de même pour les entrées adjectivales. Considérons par exemple le lexème adjectival INSUFFISANT$_1$ (le 1 ne présuppose pas qu'il existe un INSUFFISANT$_2$, qui, en l'espèce, n'existe pas dans le Le*fff*). Il s'agit d'un adjectif prédicatif qui a deux arguments : un sujet (ce qui est insuffisant) et un argument oblique en *pour*[19]. Il peut être utilisé de façon habituelle mais également au sein d'une construction impersonnelle (*il est insuffisant (pour convaincre qui que ce soit) d'affirmer que cette erreur était inévitable*). Il aura ainsi deux redistributions possibles : l'une, *%adj_personnel*, correspond aux emplois habituels, l'autre, *%adj_impersonnel*, aux cas où l'adjectif est la tête sémantique d'une construction impersonnelle. D'où l'entrée suivante :

INSUFFISANT$_1$ *adj-4* Lemma ;adj ;<*Suj* :*cln* |*scompl* |*sinf* |*sn*,
Obl :(*pour-sinf* |*pour-scompl* |*pour-sn*)> ;
cat=adj ;
%adj_personnel,*%adj_impersonnel*

19. Les arguments introduits par la préposition *pour* ont été largement sous-étudiés lors du développement des principales ressources lexico-syntaxiques pour le français. Nous avons effectué récemment un premier pas vers leur étude systématique (Sagot *et al.*, 2014).

5.1.3.3 Entrées lexicales complexes

Dans son état actuel, l'implémentation d'Alexina permet de représenter deux types d'entrées lexicales complexes, qui correspondent à certains types de mots lexicaux et en approximent d'autres. Il s'agit des deux types suivants :

1. Les lexèmes typographiquement multiples mais syntaxiquement et sémantiquement atomiques ('*words with spaces*') ; leurs propriétés syntaxiques sont identiques à celles des entrées simples ; au niveau morphologique, en revanche, ces lexèmes sont de deux sous-types :

    1a. Ils peuvent être morphologiquement atomiques, et se fléchissent alors, sauf à ne pas être fléchis, selon une classe flexionnelle unique qui fléchit l'ensemble typographique en ignorant effectivement totalement la présence d'espaces (*open space(s), parce que*) ;

    1b. Ils peuvent être morphologiquement multiples, les composants morphologiques pouvant alors se fléchir ; Alexina permet de spécifier une classe flexionnelle pour chaque token, ce qui fait l'hypothèse que les composants morphologiques fléchis coïncident avec des tokens ; ceci est suffisant dans la grande majorité des cas (*pomme(s) de terre*), mais pas toujours, soit parce que deux des tokens se fléchissent et que toutes les combinaisons ne sont pas permises, en général en raison d'un phénomène d'accord interne (*terre cuite*, *terres cuites*, mais **terre cuites* et **terres cuite*) [20]. En pratique, dans l'état actuel d'Alexina, il faut donc se restreindre aux cas où seul l'un des tokens se fléchit. Alexina confère alors à la forme fléchie dans son ensemble les propriétés morpho-syntaxiques conférées par la seule classe flexionnelle non-invariable au token qu'elle fléchit.

2. Les lexèmes typographiquement, morphologiquement et syntaxiquement multiples mais sémantiquement atomiques (prédicats complexes, locutions) ; Alexina ne gère que les cas impliquant d'une part une tête syntaxique et d'autre part un ensemble syntaxiquement atomique : *prendre le taureau par les cornes* est ainsi représenté comme étant constitué d'une séquence strictement figée *le taureau par les cornes* et une tête syntaxique *prendre*, ces deux éléments étant à combiner en syntaxe ; les propriétés syntaxiques précisant la façon dont se fait la combinaison sont déclenchées par des catégories spéciales attribuées à ce type d'entrées, le reste étant à spécifier dans la grammaire ; deux exemples tirés du français :

    2a. Les prédicats complexes de type V-N (sans déterminant), dont certaines sont des constructions à verbe support (*avoir faim*) mais pas toutes (*avoir*

20. Une modélisation de l'accord au sein du processus de flexion de ce type d'entités serait relativement simple à mettre en place, mais ne l'a pas encore été dans l'implémentation d'Alexina.

froid) ; le meilleur moyen d'encoder de telles entrées dans la version actuelle d'Alexina est de créer une entrée spéciale pour le nom prenant part à une telle construction, et de spécifier dans l'attribut spécial *lightverb* quelle est ou quelles sont les têtes syntaxiques possibles pour la construction (par exemple, *avoir*, *prendre* ou *attraper* pour le nom *froid*), charge ensuite à une grammaire de contraindre la cooccurrence entre cette tête syntaxique et le nom prédicatif dans une construction complexe [21] ; la flexion éventuelle du nom (ici, *froid*), est décrite comme pour tout autre nom ; la catégorie lexicale et syntaxique associée à une telle entrée est alors *cfi* (constituant figé inséparable, qui est tout à fait séparable de la tête syntaxique mais seulement par un ajout) [22].

2b. Les locutions verbales figées, dont les entrées intensionnelles sont complètes, comme *prendre le taureau par les cornes* ; Alexina permet d'identifier quel token (dont il est fait l'hypothèse qu'il coïncide avec un mot morphologique) est la tête syntaxique de l'entrée (ici, le premier, *prendre*) ; ce token verbal, ainsi que les éventuels clitiques qui l'entourent (*ne pas se prendre* dans *ne pas se prendre pour n'importe qui*, sont alors associés à la classe flexionnelle spéciale *0* qui les escamote des formes extensionnelles ; la tête syntaxique (soit le premier token, soit le token entouré d'accolades) est cependant capturé pour en faire la valeur d'un attribut spécial *synt_head*, au fonctionnement identique à *lightverb*, ajouté aux propriétés syntaxiques des entrées extensionnelles, la grammaire ayant ici encore pour rôle d'associer la tête syntaxique et le reste de la locution ; la flexion éventuelle des autres tokens est décrite comme pour le cas de *pomme de terre* ; ainsi, *faire usage (de)* sera représenté par une entrée dont la forme de citation est *faire usage*, dont la seule forme extensionnelle est *usage* (invariable dans cette locution) et dont les propriétés de valence stipulent une structure à deux arguments, un sujet et un objet indirect en *de*. La catégorie lexicale et syntaxique associée à une telle entrée est alors *cf* (constituant figé, séparable de la tête syntaxique par un des arguments) [23].

5.1.4 Le niveau syntaxique d'Alexina est-il adaptable à d'autres langues ?

Le Le*fff* n'est pas le seul lexique Alexina pour lequel des informations syntaxiques sont disponibles : outre d'autres ressources pour le français converties au format Le*fff* (cf. section 3.1.2), le lexique Le*ff*e de l'espagnol dispose également d'informations

21. Il est généralement considéré que le(s) verbe(s) support(s) associé à un nom prédicatif ne sont pas prédictibles. Ainsi on dit *prendre une décision*, *faire une sieste* et *poser une question*, mais en anglais *make a decision* '(lit.) faire une décision', *take a nap* '(lit.) prendre une sieste' et *ask a question* '(lit.) demander une question'. Nous avons toutefois eu l'occasion d'étudier et de discuter de cette non-prédictibilité supposée (Samvelian *et al.*, 2011).

22. Cf. *prendre avantage de N* vs. **prendre de N avantage*.

23. Cf. *avoir qqch sur la conscience* vs. *avoir sur la conscience qqch*.

syntaxiques, quoique d'un niveau de couverture et de précision bien inférieur. De plus, des expérimentations non publiées car insuffisamment avancées ont été menées sur le persan. Mais c'est surtout la correspondance presque directe entre Alexina et la norme LMF quant à la façon d'encoder l'information de valence qui invite à penser que le modèle syntaxique d'Alexina pourrait encoder l'information lexicale d'autres langues sans difficulté majeure. Nous envisageons d'ailleurs dans un proche avenir de tirer parti de ressources existantes pour l'anglais afin de doter notre lexique EnLex d'une couche syntaxique, à l'image du Le*fff*, en mettant en œuvre certaines des techniques décrites à la section 5.2.1. Certes, il ne s'agit pas là d'une langue typologiquement très différente du français au niveau lexico-syntaxique, mais on peut être confiant au vu de l'existence de ressources pour des langues comme le tchèque, l'allemand ou le persan (cf. section A.6), dont les modèles de représentation pourraient être rendus compatibles avec Alexina.

5.2 Développement de lexiques syntaxiques

5.2.1 Développement du Le*fff*

Nous avons lancé en 2003 le développement du Le*fff* à partir du constat suivant : à cette époque, il n'existait pas de lexique syntaxique pour le français qui soit librement utilisable et dont la couverture soit importante (cf. section A.6 pour un historique des travaux dans ce domaine). Pourtant, le développement, souvent coûteux, d'un lexique syntaxique n'a d'impact sur la recherche qu'à la condition qu'il soit rendu disponible à la communauté scientifique. La construction d'un tel lexique a donc été initiée au sein du projet Atoll (Inria Rocquencourt) par Lionel Clément, avec le double objectif qu'il soit adapté au TAL tout en restant linguistiquement pertinent. Dans un premier temps, le Le*fff* s'est limité à un lexique morphologique des verbes du français, acquis automatiquement et validé manuellement selon une technique originale (Clément *et al.*, 2004 ; Sagot, 2005a ; voir également les sections 3.1.1 et 3.1.2). C'est le Le*fff* 1, distribué depuis 2004.

Dans un second temps, nous avons étendu le Le*fff* à l'ensemble des catégories [24], tout en devenant un lexique morphologique *et* syntaxique. Au départ, le passage au niveau syntaxique a été effectué manuellement (Sagot *et al.*, 2006). Mais la libre distribution de ces premières versions du Le*fff*, malgré leur caractère préliminaire, a contribué à faire prendre conscience progressivement aux développeurs des autres ressources majeures pour le français (Dicovalence, le Lexique-Grammaire, Les Verbes Français [LVF] ; cf. section A.6 pour plus d'informations et des références concernant ces ressources) de la nécessité de les rendre librement accessibles, au moins à des fins de recherche. Dicovalence est ainsi rapidement passé sous licence libre, ainsi qu'une partie des tables du Lexique-Grammaire.

24. L'extension à toutes les catégories a été faite manuellement pour les catégories fermées, et à l'aide du lexique morphologique français de Multext (Veronis, 1998) pour les noms, adjectifs et adverbes — lexique dont la libre exploitation nous a été autorisée explicitement par son principal auteur

LVF a également été mis à disposition de la communauté. Ce n'est toutefois qu'en 2011 que l'intégralité des tables du Lexique-Grammaire ont été distribuées librement, après un travail important de mise en cohérence et d'explicitation des informations syntaxiques implicites (Tolone, 2011). En parallèle, le lexique TreeLex, extrait du Corpus Arboré de Paris 7, a été développé et rendu disponible librement.

C'est dans ce contexte qu'a eu lieu la suite du développement du niveau syntaxique du Le*fff* (Sagot, 2010, 2013b). Nous avons donc choisi de tirer le meilleur parti de différentes techniques automatiques mais également des autres ressources lexico-syntaxiques, au fur et à mesure de leur développement ou de leur mise à disposition avec une licence libre. En effet, cela n'aurait pas beaucoup de sens de réaliser un effort manuel considérable dès lors que des ressources de bonne qualité et de bonne couverture existent déjà. Toutefois, l'exploitation de ressources existantes, qui reposent sur des modèles et sur des choix descriptifs différents, ne va pas de soi (Sagot et Danlos, 2008). Il en résulte la nécessité d'un travail manuel de validation ou d'édition, en aval de ces techniques automatiques ou de l'exploitation d'autres ressources. C'est d'autant plus vrai pour les catégories moins bien couvertes par les ressources existantes, c'est-à-dire, en première approximation, toutes les catégories autres que les verbes simples. Nous avons formalisé dans (Sagot et Danlos, 2008) une méthodologie de fusion de lexiques syntaxiques que nous avions mise en œuvre pendant plusieurs années (Danlos *et al.*, 2006 ; Sagot et Danlos, 2007 ; Sagot et Fort, 2007 ; Danlos et Sagot, 2008 ; Sagot *et al.*, 2008, 2009b ; Sagot et Fort, 2009 ; Sagot et Danlos, 2009, 2012) sur diverses classes d'entrées lexicales (constructions pronominales, constructions impersonnelles, verbes dénominaux en *–iser/–ifier*, adverbes en *–ment*, etc.), pour tirer le meilleur parti possible des informations contenues dans ces différentes ressources, afin d'améliorer la couverture et la qualité du Le*fff*. Enfin, un travail de fusion à plus grande échelle a ensuite eu lieu, avec pour objectif d'améliorer de façon globale la qualité du lexique verbal du Le*fff* quant aux distinctions entre entrées : si l'objectif a toujours été que chaque sens d'un lemme donné donne naissance à une entrée distincte, pour peu que ce sens corresponde à des propriétés syntaxiques homogènes, ce n'est pas complètement le cas dans la dernière version standard du Le*fff*. Nous avons donc fusionné l'ensemble de Dicovalence avec les entrées verbales du Le*fff*, validé manuellement toutes les entrées pour les lemmes sur lesquels le processus de fusion a produit un résultat douteux.

Des travaux complémentaires ont également conduit au développement des informations syntaxique du Le*fff*. Par exemple, un travail manuel important a été réalisé sur l'ensemble des catégories fermées (prépositions, conjonctions…). Des techniques d'extraction automatique d'informations lexicales à partir de corpus étiquetés morphosyntaxiquement ont également été mises en œuvre, par exemple pour extraire les noms les plus fréquents impliqués dans des constructions de type V-N et leur éventuel argument prépositionnel (détecté comme étant obligatoire ou facultatif), les adjectifs susceptibles d'être en position

d'épithète antéposé (et avec quelle fréquence) ou encore des cadres de sous-catégorisation partiels pour les verbes simples, afin de combler certains manques (Sagot, 2006, ch. 7, section 2). Des expériences similaires ont également été réalisées pour identifier les verbes pouvant être la tête d'incises de citation (Sagot *et al.*, 2010 ; Danlos *et al.*, 2010) et pour les verbes sous-catégorisant un argument en *pour* (Sagot *et al.*, 2014).

Seuls les noms n'ont pas encore fait l'objet de travaux de ce type, bien que nombre d'entre eux sous-catégorisent des arguments. On peut tout d'abord penser s'appuyer sur les tables du Lexique-Grammaire, seule ressource de grande ampleur ayant abordé cette catégorie au niveau syntaxique, pour combler ce manque. Mais des travaux à l'interface entre morphologie dérivationnelle et syntaxe pourraient également y contribuer fortement (noms déverbaux, noms déadjectivaux), les propriétés de valence étant souvent conservées, au moins en partie et modulo des transformations partiellement systématiques, le long d'une relation dérivationnelle. Nous réservons cela à des travaux futurs, qui pourraient donc prendre place dans le cadre plus général de la mise en réseau des entrées du Le*fff*, réseau construit notamment à partir des relations dérivationnelles.

5.2.2 Détection automatique d'entrées morphologiques ou syntaxiques erronées

Le développement d'une ressource lexicale morphologique et syntaxique comme le Le*fff* étant un travail de longue haleine, il est inévitable que des erreurs, des manques et des incohérences s'y glissent. Mais la quantité d'informations que rassemblent de telles ressources rend difficile la détection manuelle de ces différents types d'erreurs. Certes, une entrée manquante pourra être détectée voire ajoutée au niveau morphologique grâce à des techniques comme celles présentées ci-dessus, ou plus simplement au moyen, s'ils sont disponibles, d'étiqueteurs morphosyntaxiques. Mais une entrée dont les informations syntaxiques sont incorrectes ou partielles, par exemple parce qu'un argument syntaxique est manquant ou que la liste des réalisations possibles d'une fonction syntaxique est incomplète, est bien plus difficile à détecter. C'est la ligne directrice de travaux que nous avons menés depuis 2006 (Sagot et Villemonte de La Clergerie, 2006, 2008), destinés en premier lieu à améliorer la qualité du Le*fff*, et qui reposent sur l'utilisation d'analyseurs syntaxiques symboliques ou hybrides tels que FRMG (Thomasset et Villemonte de La Clergerie, 2005 ; Villemonte de La Clergerie *et al.*, 2009a ; Villemonte de La Clergerie, 2013).

L'idée que nous avons mise en œuvre est inspirée de van Noord (2004). Pour identifier les incomplétudes et les incorrections d'un système d'analyse syntaxique, une possibilité est d'analyser un corpus de taille conséquente et étudier à l'aide d'outils statistiques ce qui différencie les phrases pour lesquelles l'analyse a réussi de celles pour lesquelles elle a échoué. L'application la plus simple de cette idée consiste à chercher les formes, dites suspectes, qui se retrouvent fréquemment dans des phrases qui n'ont pu être analysées. C'est ce que fait van Noord (2004), sans toutefois chercher à identifier (au moins) une

forme suspecte pour chaque phrase non analysable, et donc sans prendre en compte le fait qu'il y a une cause d'erreur dans toute phrase non analysable. Il définit en effet le taux de suspicion d'une forme f par le taux de phrases non analysables parmi celles contenant f.

A contrario, nous avons cherché pour chaque phrase dont l'analyse a échoué, la forme qui a le plus de chances d'être la cause de cet échec : c'est le suspect principal de la phrase (Sagot et Villemonte de La Clergerie, 2006, 2008). Il se peut que cette forme soit renseignée dans le lexique de façon incorrecte ou incomplète, qu'elle participe à des constructions non couvertes par la grammaire, ou qu'elle illustre des imperfections de la chaîne de pré-traitement utilisée en amont de l'analyseur. Si, en outre, on dispose de deux analyseurs (symboliques) distincts mais qui utilisent la même chaîne de pré-traitement et le même lexique, comme c'est le cas pour FRMG et SxLfg, on peut éliminer une partie importante des erreurs issues de l'une ou l'autre des grammaires en confrontant les résultats obtenus avec chacun des analyseurs. On ne conserve alors quasiment plus que des suspects qui correspondent à des erreurs dans le lexique, ici le Le*ffff*. C'est ainsi qu'ont été détectées, à partir du corpus journalistique du Monde Diplomatique, des erreurs telles que le manque d'un argument attribut pour le lexème *demeurer* au sens de *rester* (cf. *le problème demeure difficile*), l'oubli de la redistribution passive pour *terminer*, le manque de l'adverbe composé *tous azimuts*, et d'autres erreurs de ce type, difficiles à détecter autrement.

Cette idée peut être vue comme l'adaptation à un nouveau problème (et la simplification) de l'algorithme d'apprentissage automatique de lexiques morphologiques mentionné au chapitre 2 : plutôt que de chercher à attribuer à chaque forme un lemme parmi plusieurs lemmes hypothétiques possibles (et produits automatiquement), nous cherchons ici à attribuer un suspect à chaque phrase, parmi plusieurs suspects hypothétiques possibles qui sont ici chacun des mots de la phrase ; dans les deux cas, on itère un algorithme de point fixe qui fait un aller-retour entre le niveau local (la forme, la phrase) et le niveau global (l'ensemble des formes, l'ensemble des phrases). La tâche étudiée ici étant toutefois plus simple, l'algorithme utilisé s'en voit simplifié également. Nous n'en ferons pas ici la description, et renvoyons le lecteur intéressé à (Sagot et Villemonte de La Clergerie, 2008). On notera toutefois qu'il n'y a pas de raison, mis à part la dispersion des données, de ne considérer que les formes isolées comme causes possibles de l'échec de l'anlayse : les bigrammes de formes, les lemmes et les bigrammes de lemmes ont été utilisés également, parfois avec succès.

Cette approche a ensuite été appliquée dans des contextes différents, avec toujours pour objectif premier l'identification d'erreurs dans le Le*ffff* ou dans d'autres lexiques syntaxiques transformés au format Alexina et intégrés à l'analyseur (Tolone *et al.*, 2012 ; voir également la section 9.1.3). Ainsi, l'approche a permis l'adaptation rapide du Le*ffff* et de FRMG à un corpus spécialisé (botanique) (Role *et al.*, 2007). Elle a permis de guider

l'amélioration de la conversion automatique des tables du Lexique-Grammaire en un lexique syntaxique au format Alexina, processus évoqué à la section précédente. Elle a été adaptée pour proposer non plus un suspect pour chaque phrase inanalysable mais un gouverneur unique à chaque forme d'une phrase : on obtient alors un désambiguïsateur statistique appris de façon endogène, qui a été là aussi testé sur des corpus botaniques (Fernandez *et al.*, 2007). Enfin, ce travail a servi de point de départ à des travaux visant, au-delà de la seule détection des entrées lexicales vraisemblablement erronées, à en proposer automatiquement des corrections (Nicolas *et al.*, 2007, 2008a,b).

5.3 Premiers éléments d'évaluation du Le*fff*

Évaluer une ressource lexicale n'est pas une tâche aisée, puisqu'il est difficile de l'évaluer par rapport à une ressource de référence : si on disposait d'une telle ressource il ne serait pas nécessaire de développer la sienne (sauf bien entendu si une telle ressource existe mais n'est pas librement utilisable, modifiable ou redistribuable). On peut toutefois citer trois types d'évaluations :

1. évaluation comparative avec d'autres ressources comparables ;
2. évaluation manuelle de la précision voire de la couverture ;
3. évaluation orientée-tâche.

Un autre type d'évaluation, le calcul du rappel sur un corpus arboré, nécessite une mise en correspondance délicate entre les informations représentées dans le lexique et celles modélisées dans le corpus arboré, notamment dès lors qu'il n'y a pas cohérence entre les façons de distinguer arguments et modifieurs [25]. Nous ne nous attarderons pas ici sur ce type d'évaluation. Dans la suite de cette partie, nous nous contenterons de présenter rapidement les résultats d'une comparaison quantitative avec d'autres ressources (type 1). Plusieurs évaluations orientées-tâches (type 3) seront présentées au cours de la partie IV, y compris des comparaisons avec des principales autres ressources dans le contexte de l'analyse syntaxique automatique.

Le tableau 5.1 constitue une comparaison directe avec d'autres ressources lexicales en termes de couverture morphologique (nombre de lemmes distincts). On notera toutefois que l'inclusion sans fin de mots trop rares ou archaïques n'est pas nécessairement une bonne chose, surtout si une conséquence en est d'augmenter la taille du lexique et l'ambiguïté lexicale, sans amélioration significative de la couverture du lexique sur de textes réels.

25. Même une évaluation orientée-tâche dans un analyseur syntaxique peut souffrir de ce type de problème. Par exemple, pour le français, la prise en compte du travail sur la sous-catégorisation en *pour* devrait conduire un analyseur à faire de certains dépendants verbaux en *pour* un argument et non plus un modifieur. Mais dans le French TreeBank quasiment tous ces arguments sont annotés comme des modifieurs. L'amélioration de la qualité des analyses conduira donc à dégrader les résultats de l'évaluation automatique de l'analyseur et donc également la mesure orientée-tâche de la qualité du lexique.

Catégorie	**Le*fff***	**Morphalou** [26] (Romary *et al.*, 2004)	**Multext** (Veronis, 1998)	**Dicolavence** (van den Eynde et Mertens, 2006)
verbes	6 825	8 789	4 782	3 729
noms	37 530	59 002	18 495	0
adjectifs	10 483	22 739	5 934	0
adverbes	3 584	1 579	1 044	0
prépositions	225	(51)	117	0

TABLEAU 5.1 – Comparaison quantitiative du nombre de lemmes distincts dans différentes ressources lexicales pour le français.

5.4 Perspectives

La perspective la plus directe quant au développement du Le*fff* est naturellement la poursuite du renseignement des informations syntaxiques pour les catégories non verbales. Comme on peut le comprendre en creux à la lecture de ce chapitre, seuls les lexèmes verbaux simples ont véritablement atteint un degré de maturité satisfaisant, qui font du Le*fff* une des ressources de référence du français, et probablement le lexique syntaxique le plus utilisé en traitement automatique des langues pour cette langue. Mais seules les entrées adjectivales ayant été l'objet de l'étude sur les constructions impersonnelles et de celle sur les arguments en *pour*, ainsi que celles concernant des adjectifs présents dans le French TreeBank et donc dans TreeLex disposent d'informations syntaxiques précises. Concernant les noms, seuls les plus fréquents parmi ceux prenant part à des constructions de type V-N sont correctement couverts. Quant aux locutions verbales figées, seules certaines classes ont été intégrées à ce jour. Il y a donc encore une marge de progression importante. On peut noter toutefois que l'annotation des arguments nominaux et adjectivaux dans le French TreeBank, tout comme celle des locutions verbales figées et des constructions à verbe support, est très perfectible. Une meilleure description des catégories autres que les verbes simples dans le Le*fff* devrait donc améliorer la qualité d'un analyseur comme FRMG qui repose sur ce lexique, mais il n'est pas certain que cela se traduise par des résultats quantitatifs significatifs, voire simplement positifs, dans les résultats de l'application de mesures classiques d'évaluation des analyseurs syntaxiques par comparaison avec le French TreeBank. Nous avons déjà évoqué cette difficulté à propos des arguments en *pour*.

Avant de passer au prochain chapitre au niveau sémantique, et en guise de transition, il est loisible de rappeler que les lexiques Alexina, et notamment le Le*fff*, sont en un certain sens des lexiques sémantiques, puisque l'inventaire d'entrées lexicales,

26. On notera que dans Morphalou les variantes masculines et féminines des noms tels que boulanger/boulangère y forment deux lemmes distincts, alors qu'elles sont considérées (de façon discutable) comme des formes d'un même lemme dans le Le*fff*. C'est une des causes du nombre élevé de lemmes nominaux dans Morphalou.

supposées correspondre à des lexèmes, est déterminé par des propriétés morphologiques, syntaxiques et sémantiques. C'est la raison pour laquelle une des perspectives d'évolution pour le Le*fff* est d'associer chacune de ses entrées à des identifiants sémantiques tels que disponibles dans le lexique sémantique WOLF, le wordnet libre du français développé par des techniques automatiques auxquelles le prochain chapitre est consacré. Ce couplage pourrait d'ailleurs se faire avec ou grâce à deux autres ressources en cours de développement, à savoir le lexique partiel à la FrameNet développé pour le français dans le cadre du projet ASFALDA (Candito *et al.*, 2014) d'une part et le VerbNet français d'autre part (Verbɔnet, Pradet *et al.*, 2014b).

Une autre évolution, elle aussi à l'interface avec la sémantique, est l'intégration de liens de dérivation morphosémantiques. Ce travail, dont les préliminaires purement morphologiques ont déjà été évoqués au chapitre 2, pourrait ouvrir de nouvelles perspectives de recherche quant au traitement des mots inconnus, et notamment des néologismes, mais également en vue d'études morphologiques quantitatives.

Enfin, il serait intéressant d'appliquer le même type de techniques de développement de lexiques syntaxiques à d'autres langues que le français. Les premières expériences dans cette direction, réalisées sur l'espagnol (Molinero *et al.*, 2009b), n'ont pas avancé depuis plusieurs années, faute de temps et de collaborateurs. Mais des ressources libres existent désormais pour des langues comme l'espagnol, le catalan, l'allemand, le tchèque ou encore l'anglais. Enrichir les lexiques morphologiques Alexina pour ces langues d'informations syntaxiques pertinentes ouvrirait la voie à de nombreux travaux, comme par exemple le développement d'une métagrammaire multilingue à partir de la métagrammaire FRMG (Villemonte de La Clergerie, 2013), en factorisant au mieux les éléments communs à plusieurs langues et en ne spécifiant que les différences. Couplées à des lexiques syntaxiques Alexina, de telles métagrammaires permettraient le développement d'analyeurs syntaxiques très performants, à l'image de FRMG pour d'autres langues que le français.

De plus, si le couplage entre niveaux syntaxique et sémantique est réalisé, disposer de ressources lexico-syntaxiques dans d'autres langues permettrait la mise en correspondance d'entrées lexicales syntaxiques dans plusieurs langues, à la fois au niveau des lexèmes (relation de traduction) et au niveau des cadres de sous-catégorisation. Il serait alors intéressant d'étudier l'utilité de telles ressources dans des contextes telles que l'extraction d'informations multilingue ou la traduction automatique.

Chapitre 6

Développement du WOLF et de sloWNet, wordnets libres du français et du slovène [1]

Sommaire

Au cours de la dernière décennie, le rôle des ressources lexicales de niveau sémantique ou encyclopédique s'est considérablement accru au sein du domaine du traitement automatique des langues, divers travaux montrant l'intérêt de telles ressources pour améliorer les performances pour divers types de tâches. Par exemple, Gabrilovich et Markovitch (2006) ont prouvé que l'utilisation de connaissances encyclopédiques améliore la classification automatique de documents. De même, Nastase (2008) a mis en œuvre de telles connaissances pour améliorer le résumé automatique. Harabagiu *et al.* (2000) ont obtenu des améliorations dans un système de réponse à des questions en tirant parti des liens entre mots dans un réseau lexical sémantique, dont l'intérêt a également été montré pour des tâches comme la désambiguïsation lexicale (Cuadros et Rigau, 2006) ou la traduction automatique (Carpuat et Wu, 2007).

Un certain nombre d'architectures ont été proposées pour organiser et représenter les connaissances lexicales sémantiques, telles qu'ACQUILEX [2] (Copestake *et al.*, 1993), le

1. Le travail a été réalisé en partie en collaboration avec Darja Fišer (Université de Ljubljana), y compris dans le cadre d'un projet bilatéral PROTEUS franco-slovène dont j'étais le responsable pour la partie française.

2. http ://www.cl.cam.ac.uk/research/nl/acquilex/ [06.07.2014]

Roget's Thesaurus [3] (Kirkpatrick, 1987), MindNet [4] (Richardson *et al.*, 1998), ConceptNet [5] (Liu, 2003) ou Cyc [6] (Matuszek *et al.*, 2006). D'autres types de ressources, et notamment celles produites par les projets de la famille FrameNet (Baker *et al.*, 1998), s'intéressent plus spécifiquement à la valence sémantique des prédicats, et relèvent ainsi plutôt de la sémantique prédicative, à mi-chemin avec les problématiques évoquées au chapitre 5 sur les lexiques dits syntaxiques [7].

Mais l'une des ressources sémantiques lexicales les plus connues et les plus utilisées dans les domaines du traitement automatique des langues et du web sémantique est le Princeton WordNet (PWN) (Fellbaum, 1998) et ses équivalents pour d'autres langues, notamment les wordnets développés dans le cadre des projets EuroWordNet (Vossen, 1999), BalkaNet (Tufiş, 2000) ou AsianWordnet (Sornlertlamvanich, 2010), ainsi que le récent Open Multilingual Wordnet [8], qui normalise et fusionne tous les wordnets dont la redistribution est autorisée par leurs auteurs, et inclut à ce jour des wordnets pour 27 langues. Initialement, le PWN était pourtant développé dans un contexte psycho-lexicographique (Miller, 1995), inspirée par des travaux sur les processus cognitifs d'accès au lexique.

Dans un wordnet, les lexèmes sont organisés en ensembles de synonymes, ou synsets, chaque synset représentant un sens. Un synset a un identifiant unique et contient donc un certain nombre de littéraux, qui sont approximativement des lemmes (simples ou composés), des termes voire des collocations, qui tous peuvent exprimer le sens représenté par le synset. Les synsets sont reliés entre eux par des relations sémantiques, la plus structurante étant la relation d'hypéronymie. Parmi les autres relations incluses dans le PWN on peut citer les relations de méronymie, d'holonymie ou d'antonymie. Par exemple, dans la version 3.1 du PWN, le synset nominal d'identifiant 02086723-n contient les littéraux {*dog, domestic dog, Canis familiaris*}. Le sens ainsi représenté est illustré par

3. http ://www.bartleby.com/62/ [06.07.2014]
4. http ://research.microsoft.com/nlp/Projects/MindNet.aspx [06.07.2014]
5. http ://web.media.mit.edu/ hugo/conceptnet/ [06.07.2014]
6. http ://www.cyc.com/ [06.07.2014]
7. FrameNet est un projet toujours actif, démarré en 1997, qui s'appuie sur la sémantique des *frames*, développée par Fillmore (1968, 2006, 1982). Dans cette théorie, les mots n'ont de sens qu'en référence à un espace conceptuel structuré qui, dans le projet FrameNet, est mis en œuvre par des liens entre les *frames*, chacun d'entre eux étant une structure conceptuelle « qui décrit un type particulier de situation, d'objet ou d'événement, ainsi que ses participants (actants) et ses propriétés » (traduction libre de (Ruppenhofer *et al.*, 2005)). Ces liens hiérarchisent l'ensemble des *frames* mais sont de différentes natures. Il est important de noter que les participants aux *frames*, ou *frame elements*, sont déterminés (en principe) au moyen de critères purement sémantiques, sans référence aux cadres argumentaux des lexicalisations de ces *frames*. Le FrameNet de l'anglais a été développé avec une approche très lexicographique, le développement de l'inventaire de *frames* et leurs lexicalisations possibles ayant fortement guidé l'annotation de quelques phrases pour chaque *frame*, seule la lexicalisation du *frame* concerné étant alors annotée. L'annotation de texte tout-venant n'en est qu'à ses débuts, et pose des problèmes importants. Depuis, d'autres FrameNet ont vu le jour, notamment pour l'espagnol, l'allemand, le suédois, et désormais le français, dans le cadre du projet ANR ASFALDA porté par Marie Candito et auquel nous avons participé (Candito *et al.*, 2014). Dans la majorité des cas, une approche plus équilibrée entre annotation de corpus et développement du lexique a été employée.
8. http ://compling.hss.ntu.edu.sg/omw/ [06.07.2014]

une définition (*a member of the genus Canis [...] that has been domesticated by man since prehistoric times ; occurs in many breeds*) et un exemple d'emploi (*the dog barked all night* 'le chien a aboyé toute la nuit'). Ce synset a deux hypéronymes, les synsets 02085998-n {*canine, canid*} 'canidé' et 01320032-n {*domestic animal, domesticated animal*} 'animal domestique'. Il a un certain nombre d'hyponymes, dont par exemple les synsets 02089774-n {*hunting dog*} 'chien de chasse' et 02113929-n {*Newfoundland, Newfoundland dog*} 'terre-neuve' [9].

Les premiers wordnets, et notamment le PWN, ont été développés manuellement, afin de maximiser la pertinence linguistique et de minimiser le taux d'erreur. Cependant, pour la grande majorité des langues, un tel effort est bien trop coûteux en temps et en moyens humains pour pouvoir être reproduit. C'est la raison pour laquelle diverses approches semi-automatiques et totalement automatiques ont été proposées pour le développement de wordnet à partir de divers types de ressources, et notamment en s'appuyant sur la disponibilité préalable du PWN. Ces approches diffèrent à plusieurs niveaux : équilibre recherché entre précision et couverture (plus la couverture visée est grande, plus le taux d'erreur dans la ressource produite est élevé), degré de complexité des ressources utilisées (depuis de « simples » lexiques bilingues jusqu'à des thésaurus complexes, depuis des corpus monolingues bruts jusqu'à des corpus multilingues parallèles), degré de complexité des algorithmes employés, etc. On pourra se rapporter à la section A.7 pour un aperçu des travaux antérieurs sur le développement semi-automatiques et automatique de wordnets.

Nous présentons dans ce chapitre la méthodologie générale de développement de wordnets que nous avons mise en place et appliquée au développement de wordnets pour deux langues : le français et le slovène. Il s'agit ainsi d'une méthodologie indépendante de la langue. Elle s'appuie sur la disponibilité de ressources libres, et notamment de corpus parallèles et de ressources de type wiki (Wikipedia et Wiktionary/Wiktionnaire), pour « traduire » en français et en slovène le Princeton Wordnet (PWN), wordnet libre de l'anglais (Fellbaum, 1998). Pour une description du modèle wordnet et un historique du développement de wordnets, d'abord manuel pour l'anglais avec une ambition psycho-lexicographique (Miller, 1995), puis pour d'autres langues, manuellement ou automatiquement [10], on pourra se reporter à la section A.7.

Il est peut-être utile de rappeler la pertinence du développement de nouveaux wordnets pour le français et le slovène. Au moment du démarrage des travaux rapportés dans ce chapitre (début 2008), le seul wordnet qui existait pour le français avait été développé dans le cadre du projet EuroWordNet (Vossen, 1999). Dans la suite de ce chapitre, nous

9. Sauf mention explicite du contraire, les identifiants de synsets mentionnés ici correspondent à la version 3.1 du PWN.

10. On peut citer les projets EuroWordNet (Vossen, 1999), BalkaNet (Tufiş, 2000) ou AsianWordnet (Sornlertlamvanich, 2010), ainsi que le Open Multilingual Wordnet [11], qui normalise et fusionne tous les wordnets dont la redistribution est autorisée par leurs auteurs, et inclut à ce jour des wordnets pour 27 langues.

dénoterons ce wordnet par le terme French WordNet (FWN). Ceci dit, l'utilisation de cette ressource n'a jamais été très répandue, principalement pour des raisons liées à sa disponibilité et à la licence qui lui était associée. De plus, il n'y a pas eu de suite en France au projet EuroWordNet, qui aurait pu travailler à l'extension et l'amélioration de cette ressource restreinte à un sous-ensemble des noms et des verbes, à l'exclusion des adjectifs et des adverbes (Jacquin *et al.*, 2007). Depuis la création du WOLF, un autre wordnet a été développé en parallèle au moyen de ressources bilingues librement disponibles extraites de ressources wiki ainsi que d'un modèle de langue syntaxique du français. La première version, limitée aux synsets nominaux, est distribuée sous le nom de JAWS (Mouton et de Chalendar, 2010). Une version ultérieure, obtenue grâce à une version améliorée de la méthode et couvrant toutes les parties du discours, est distribuée sous le nom de WoNeF (Pradet *et al.*, 2014a) et évaluée avec soin. Nous comparerons donc le WOLF avec le FWN, JAWS et WoNeF. Pour le slovène, un wordnet de taille très modeste avait déjà été développé, en partie grâce à l'utilisation du wordnet serbe développé dans le cadre du projet BalkaNet et de la proximité entre le serbe et le slovène (Erjavec et Fišer, 2006 ; Fišer, 2007). Ce Slovene WordNet, malgré le petit nombre de synsets qu'il contient, reste néanmoins utile, comme nous le verrons, pour servir de base à des évaluations quantitatives, puisqu'il a été entièrement validé manuellement.

Nous décrirons brièvement les sources d'informations lexicales que nous avons utilisées et la façon dont nous avons extrait ces informations. Nous donnerons ensuite un aperçu des différentes techniques employées pour la construction du WOLF et du sloWNet au fil des années. Nous proposerons enfin une évaluation succincte de la qualité de ces ressources. Pour une description plus détaillée de nos travaux, on pourra consulter notamment (Sagot, 2017c), qui se concentre sur le WOLF, et (Fišer et Sagot, 2015), qui se concentre sur sloWNet, ainsi que les différents articles ayant décrit ces travaux au fil des années et cités au fil du ce chapitre.

6.1 Extraction automatique d'informations lexicales sémantiques

Un des points de départ de la plupart des techniques de développement automatique de wordnets est la disponibilité de ressources lexicales sémantiques bilingues ou multilingues. L'idée est tout d'abord d'extraire de telles ressources des couples de type *(littéral anglais, littéral français/slovène)* puis, par rapprochement avec le PWN, de

construire le plus grand nombre possible de couples *(littéral, synset)* candidats dans la langue cible. Reste ensuite à sélectionner les meilleurs d'entre eux [12, 13].

Les ressources desquelles on extrait des lexiques bilingues ou multilingues peuvent être de trois types :

1. Des dictionnaires bilingues ou multilingues, qui ne fournissent dans le cas général que des paires de littéraux en relation de traduction. Nous avons utilisé notamment des ressources de type Wiktionnaire (le Wiktionary anglais et les Wiktionnaires français et slovènes, notamment) [14]
2. Des vedettes d'articles de l'encyclopédie collaborative en ligne Wikipedia, disponible pour de nombreuses langues avec des liens inter-langues qui relient des (vedettes d') articles dénotant généralement le même concept ou la même entité [15]
3. Des corpus parallèles multilingues. Nous avons utilisé le corpus SEE-ERA.NET, un sous-corpus d'environ 1,5 million de mots du JRC-Acquis (Tufiş *et al.*, 2009) disponible en 8 langues dont l'anglais, le français et le slovène. Outre ces trois langues, nous avons également utilisé les données en roumain, tchèque et bulgare, qui sont les langues pour lesquelles nous disposions de wordnets alignés avec le PWN, tous issus du projet BalkaNet (Tufiş *et al.* 2009 ; cf. section A.7). Nous en avons extrait par alignement automatique des lexiques bilingues pour chaque paire de langue disponible, dans lesquels chaque entrée est associée à une mesure de confiance issue des indices de confiance de chacune de ses occurrences. [16]

12. Par exemple, en partant d'un littéral anglais du PWN qui apparaît dans n_s synsets et pour qui l'on a extrait n_t traductions, l'objectif est d'identifier les meilleurs des $n_s n_t$ candidats possibles en désambiguïsant chaque traduction en contexte, grâce à des corpus parallèles, ou hors contexte, notamment en exploitant une version précédente du wordnet dans la langue cible, ici le français ou le slovène. Naturellement, pour un littéral monosémique dans la langue source, ici l'anglais, il n'y a pas besoin de désambiguïsation : tous les candidats obtenus sont en principe valides, et l'on peut associer sans hésitation toute traduction en français ou en slovène à l'unique synset auquel il appartient dans le PWN.

13. Dans ce chapitre, nous qualifions de « monosémique » un littéral anglais dès lors qu'il n'apparaît que dans un seul synset du PWN. C'est naturellement une approximation, qui repose sur l'hypothèse selon laquelle le PWN est complet. Cette approximation se justifie d'une part par la grande couverture de cette ressource et d'autre part par le fait qu'un littéral anglais qui est en réalité polysémique, s'il n'apparaît que dans un seul synset du PWN, a de fortes chances d'avoir ce sens unique pour sens très majoritaire : ainsi, il est plausible que les traductions que nous pourrions en trouver en français ou en slovène concernent toutes, ou presque toutes, ce sens majoritaire.

14. Nous avons extrait du Wiktionary anglais et du Wiktionnaire français respectivement 62 826 et 59 659 entrées bilingues anglais–français. Du Wiktionary slovène et du Wiktionary slovène nous avons extrait 6 052 et 7 029 entrées bilingues anglais–slovène, soit des nombes bien inférieurs. Nous avons eu heureusement accès à des dictionnaires bilingues anglais–slovène (Grad *et al.*, 1999) et slovène–anglais (Grad et Leeming, 1999) en complément, dont nous avons pu extraire respectivement 207 972 et 72 954 entrées bilingues anglais–slovène. Nous avons également extrait des entrées bilingues d'autres types de ressources, pour lesquelles nous renvoyons à (Fišer et Sagot, 2015 ; Sagot, 2017c).

15. Nous avons ainsi extrait 286 822 couples *(littéral anglais, littéral français)* et de 32 669 couples *(littéral anglais, littéral slovène)* qui sont toutes considérés comme nominaux (noms propres ou noms communs, qui ne sont pas différenciés dans le PWN.

16. Nous avons utilisé différents outils pour étiqueter morpho-syntaxiquement et lemmatiser ces textes, avant de les aligner au niveau des phrases puis des mots à l'aide de l'outil Uplug (Tiedemann, 2003).

La création ou l'extension d'un wordnet au moyen d'une approche qui préserve le même inventaire de synsets que le PWN peut être vue comme une tâche consistant à produire des couples *(littéral en langue cible, synset)*, que nous appelons des *candidats*, par l'intermédiaire d'une entrée bilingue *(littéral en langue cible, littéral en anglais)*.

Lorsque l'on part d'entrées extraites de ressources telles que le Wiktionaire ou Wikipedia, il faut donc identifier, parmi les synsets possibles dans le PWN pour le littéral en anglais, le synset qui correspond au sens qui peut être exprimé dans la langue cible par le littéral en langue cible associé. Si le littéral anglais est monosémique (dans le PWN), la tâche est triviale [17] Dans les autres cas, l'ensemble des candidats possibles est bien plus bruité, ce qui nécessite le développement d'un mécanisme dédié de désambiguïsation [18, 19].

Lorsque l'on part d'entrées extraites de corpus alignés, on dispose de la liste de leurs occurrences en corpus, au sein du corpus multilingue de départ. Ceci permet de reconstituer des entrées non plus bilingues mais multilingues, par exemple en extrayant les équivalents de traduction dans toutes les autres langues de chaque occurrence de

L'alignement ayant été réalisé au niveau des mots simples, nous n'avons pas pu, par cette méthode, extraire d'entrées bilingues comportant des mots ou des termes composés. Le lexique le plus petit que nous ayons extrait, après un seuillage sur les mesures de confiance et les nombres d'occurrence de chaque lien, est le lexique tchèque–anglais (43 024 entrées), le plus gros étant le lexique tchèque–bulgare (50 289 entrées).

17. Par exemple, le littéral anglais *battle of Gettysburg* est monosémique, puisqu'il n'apparaît que dans un seul synset, le synset 01282108-n. Or nous avons extrait de Wikipedia la traduction française *bataille de Gettysburg* pour ce littéral. Nous pouvons donc produire le candidat français (*bataille de Gettysburg*, 01282108-n). On peut remarquer, à propos de cet exemple, deux des limites du modèle wordnet et de l'utilisation de l'inventaire de synsets du PWN pour développer des wordnets pour d'autres langues. La première limite est le niveau de granularité atteint par cet inventaire de synsets. Le second, qui lui est indirectement lié, est le biais culturel, géographique et civilisationnel qui est parfois manifeste dans le PWN, ressource développée par des américains pour modéliser l'anglais américain (Orav et Vider, 2004 ; Wong, 2004). Par exemple, le PWN contient un synset {*performer*, *performing artist*}, défini comme étant un artiste réalisant un spectacle théâtral ou musical devant face à une audience. Mais il n'y a pas de mot en français qui dénote de façon globale les acteurs, chanteurs et autres artistes se produisant en spectacle. Dans un tel cas, il est toujours possible de laisser vide le synset en question dans la ressource produite. À l'inverse, certains sens raisonnablement répandus de la langue cible peuvent ne pas correspondre à un synset du PWN, notamment lorsqu'ils n'ont pas vraiment de réalité culturelle aux États-Unis ou ne sont pas considérés comme suffisamment importants. C'est ainsi le cas des sens ou concepts dénotés en français par {*raclette*} ou ou {*École Polytechnique*}. Parfois, c'est le découpage même en synsets qui ne correspond pas bien. Ainsi, les synsets {*lawyer*, *attorney*} (« a professional person authorized to practice law ; conducts lawsuits or gives legal advice ») et {*advocate*, *counsel*, *counselor*, *counsellor*, *counselor-at-law*, *pleader*} (« a lawyer who pleads cases in court ») sont distingués selon des critères propres au système judiciaire américain, qui ne se superposent pas avec les distinctions françaises entre juriste, avocat et avoué. Il n'en reste pas moins que les bénéfices du développement de worndets alignés sur le PWN restent supérieurs à ces réels défauts.

18. Par exemple, le littéral anglais *dog* est dans huit synsets. Le lien de traduction anglais–français *(dog, chien)* donnera donc lieu à autant de candidats français impliquant le littéral français *chien*, dont la plupart sont vraisemblablement erronés.

19. Dans certains articles du Wiktionary anglais (et dans d'autres langues), les traductions d'un mot donné sont parfois triées en sens, chacun étant associé à une très courte glose. Certains auteurs ont proposé de comparer ces gloses à celles du PWN pour induire des liens entre sens du Wiktionary et synsets (Bernhard et Gurevych, 2009 ; Casses, 2010). La première phrase d'un article de Wikipedia peut être utilisée de la même façon (Ruiz-Casado *et al.*, 2005). Toutefois, c'est loin d'être le cas de toutes les entrées de Wiktionary. Et lorsque c'est le cas, les gloses sont souvent réduites à un mot ou à quelques mots tout au plus. Nous avons donc décidé de ne pas utiliser ces informations, et d'intégrer les entrées bilingues dans notre processus plus global de désambiguïsation.

Tchèque	**Bulgare**	**Anglais**	**Français**
právo	право	law	droit
06129345	04893549	00577416	→*05791721*
05559593	04888072	05529208	
05791721	07928837	05531141	
04617988	00577416	**05791721**	
07928837	**05791721**	06129345	
	01000872	07712371	
	04881053	07928837	
	04617988		

Tableau 6.1 – Illustration du processus de désambiguïsation des entrées multilingues extraites de corpus alignés. Les identifiants de synsets proviennent de la version 2.0 du PWN, conformément aux versions d'origine des wordnets BalkaNet. Les exemples étant tous nominaux, un numéro de synset comme 06129345 est donc ici à interpréter comme ENG20-06129345-n.

chaque mot de la langue cible [20]. Naturellement, les erreurs d'alignement conduisent à ce que ces lexiques multilingues sont partiellement bruités. Ces lexiques permettent alors de réaliser une désambiguïsation sémantique par intersection de la façon suivante. Pour chaque entrée multilingue, nous avons extrait des wordnet BalkaNet, alignés sur le PWN, l'ensemble des synsets associés à chaque littéral de chaque langue concernée (cf. tableau 6.1). On peut alors considérer, si l'entrée est valide, que le sens véhiculé par les différents littéraux qui composent cette entrée, lorsqu'ils sont en lien de traduction, est représenté par l'intersection des synsets contenant chacun des littéraux. [21]

6.2 Développement du WOLF et de sloWNet

Les versions actuelles du WOLF et de sloWNet sont le résultat de divers travaux, que nous ne détaillerons pas tous ici mais qui sont récapitulés à la figure 6.1. Ils se structurent autour de trois étapes principales, celles dont l'impact quantitatif a été le plus important.

20. Nous avons ainsi extrait un ensemble de cinq lexiques multilingues qui contiennent, pour le français, entre 49 356 entrées (lexique français–roumain–tchèque–bulgare–anglais) et 59 019 entrées (français–tchèque–bulgare–anglais), et pour le slovène entre 52 193 entrées (lexique slovène-tchèque-anglais) et 55 768 entrées (slovène-tchèque-bulgare-anglais).

21. Sur l'exemple du tableau 6.1, il n'y a qu'un seul synset qui est commun au tchèque *právo*, au bulgare *право* et à l'anglais *law* : ce synset représente donc le sens véhiculé par ces trois mots ainsi que par le mot français *droit* dans toutes les occurrences où ils sont en relation de traduction quadrilingue. On peut donc proposer un candidat associant le littéral français *droit* à cet unique synset. Dans ce cas, le candidat proposé est correct, puisqu'il associe droit à un synset dont la glose dans le PWN est *the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do* 'branche de la philosophie qui s'intéresse à la loi et aux principes qui conduisent les tribunaux à prendre les décisions qu'ils prennent'. Nous avons ainsi produit plusieurs milliers de candidats à partir de chacun des cinq lexiques multilingues. Ces ensembles de candidats sont bruités, en conséquence des erreurs d'alignement signalées plus haut mais aussi des ambiguïtés résiduelles (par exemple si l'intersection entre ensembles de synsets ne donne pas un seul synset, comme dans l'exemple du tableau 6.1, mais deux).

Des informations quantiatives sur les différentes versions successives du WOLF et de sloWNet sont fournies au tableau 6.2, ainsi que des éléments de comparaison concernant le wordnet français du projet EuroWordNet (Vossen, 1999), ou French Wordnet (FWN), le wordnet nominal JAWS développé en 2010 (Mouton et de Chalendar, 2010) et son successeur WoNeF Pradet *et al.* (2014a) [22]. Les trois étapes principales, qui sont indiquées sur fond grisé dans la figure 6.1, sont les suivantes :

1. **Création des versions initiales des deux ressources** (Fišer et Sagot, 2008 ; Sagot et Fišer, 2008), dont les résultats sont le WOLF 0.1.4 et sloWNet 2.0.
   Cette première étape repose sur tous les candidats construits à partir des corpus alignés, mais en ne conservant parmi les candidats construits à partir des ressources lexicales que ceux issus d'un littéral anglais monosémique. Nous avons fusionné les candidats retenus, en gardant trace des ressources dont ils avaient été extraits et en appliquant quelques filtres heuristiques pour éliminer les candidats les moins plausibles [23]. Les candidats finalement conservés dans l'une ou l'autre des deux langues cibles, le français et l'anglais, sont alors utilisés pour peupler une version du PWN dont tous les littéraux anglais avaient été supprimés. Le résultat est un wordnet dans la langue cible qui a le même nombre de synsets que le PWN, et qui en a conservé gloses et exemples d'usage (en anglais), ainsi que tous les liens sémantiques entre synsets [24]. Certains synsets contiennent donc désormais des littéraux de la langue cible (français pour le WOLF, slovène pour sloWNet), et d'autres sont vides, mais toutes les informations structurelles ont été préservées.

2. **Extension à grande échelle**, en s'appuyant sur l'existence de versions initiales et en augmentant très significativement leur couverture (Sagot et Fišer, 2011, 2012a, 2014).
   Les résultats sont le WOLF 0.2 et sloWNet 3.0. L'objectif de ce travail était d'exploiter l'ensemble des candidats que l'on peut construire à partir des ressources lexicales, et pas seulement ceux impliquant un littéral anglais monosémique, au moyen d'une stratégie de désambiguïsation dédiée et en tirant parti de l'existence préalable de wordnets pour les langues cibles, même préliminaires. On peut en effet représenter un wordnet par les littéraux qu'il contient et ceux que contiennent des synsets qui lui sont proches au sein du PWN, puis utiliser la distribution de ces littéraux en

22. Pour WoNeF (version 0.1), nous donnons les résultats pour la version standard de cette ressource, qui en optimise le f-score (cf. `http://wonef.fr/`). Pradet *et al.* (2014a) proposent également une version optimisant la couverture, mais de moindre précision, et une version optimisant la précision, mais de bien moindre couverture.

23. Par exemple, sont éliminés les candidats uniquement proposés par un corpus aligné trilingue et avec un nombre d'occurrences très bas.

24. Nous avons également conservé les informations indiquant l'appartenance ou non de chaque synset à l'ensemble des Basic Concept Sets (BCS) définis au sein du projet BalkaNet (les synsets BCS1 sont supposés être les plus fondamentaux, les BCS2 le sont un peu moins, les BCS3 encore moins, les synsets hors BCS constituant la grande majorité d'un wordnet).

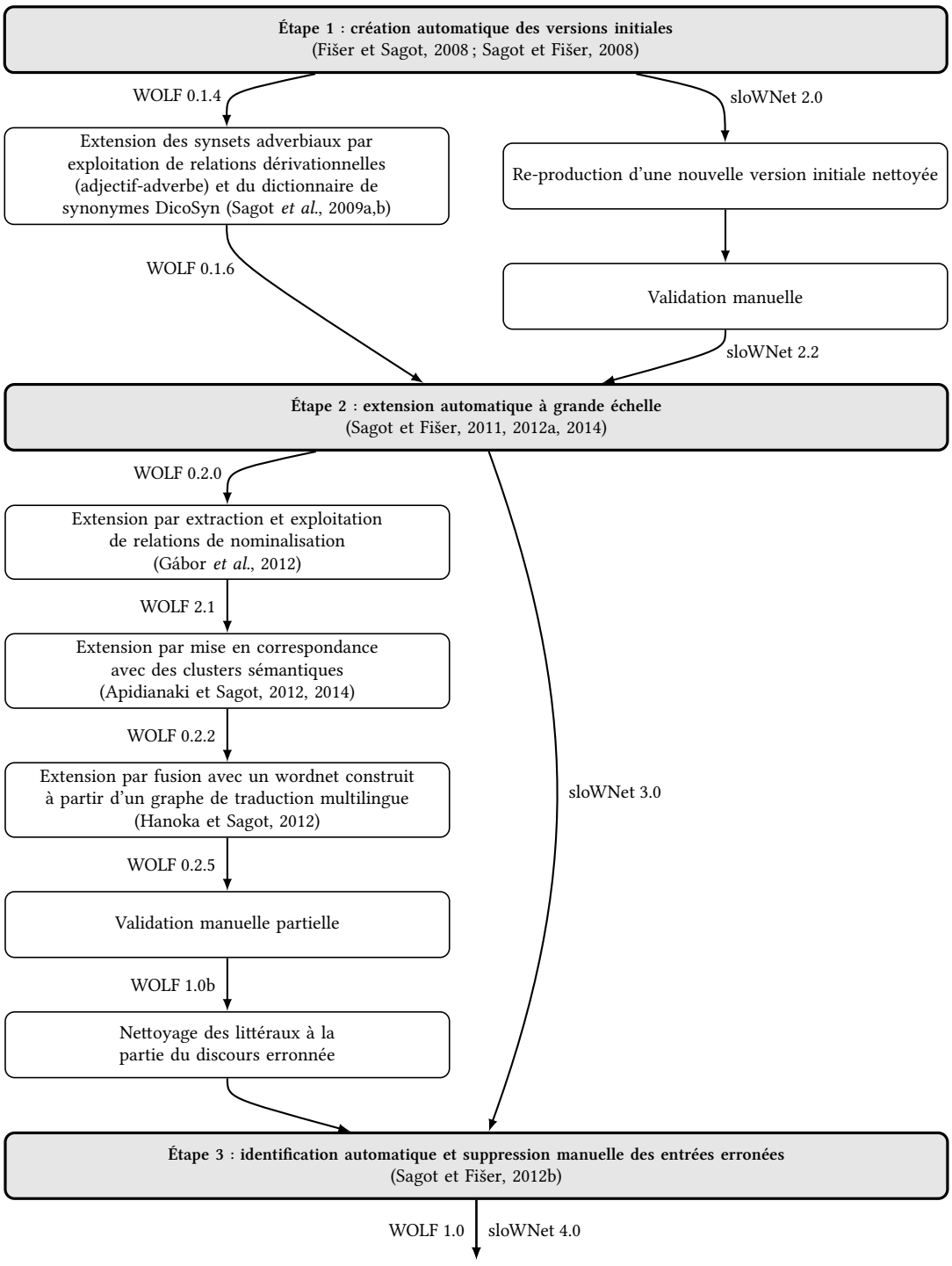


FIGURE 6.1 – Séquence de travaux mis en œuvre pour le développement de WOLF et de sloWNet. Les trois étapes majeures, en gras et sur fond gris clair, sont décrites dans la suite de ce chapitre.

corpus comme une approximation de la distribution en corpus du sens porté par le synset de départ. Nous entraînons alors un classifieur, qui utilise différents traits associés à chaque candidat *(littéral, synset)*, parmi lesquels cette mesure de similarité distributionnelle entre le littéral anglais du candidat le wordnet testé, et dont le but est de déterminer si le candidat est valide ou non. Nous donnons au classifieur retenu les données d'entraînement suivantes : parmi tous les candidats que nous avons construits, tous ceux qui sont déjà présents dans la version alors disponible du WOLF ou de sloWNet sont considérés comme des exemples positifs, et tous les autres candidats comme des exemples négatifs [25]. Nous examinons alors les résultats du classifieur sur ses propres données d'entraînement : chaque candidat reçoit alors un score compris entre 0 (candidat certainement erroné) et 1 (candidat certainement valide). Nous avons fixé empiriquement un seuil identique de 0,1 pour les deux langues : tout candidat dont le score est supérieur ou égal à ce seuil est conservé. Malgré l'utilisation hétérodoxe du concept de classifieur dans cette approche, les résultats que nous avons obtenus en montrent la validité, tant en termes de qualité des candidats [26] qu'en terme d'augmentation de la couverture des wordnets [27].

3. **Suggestion automatique et filtrage manuel d'intrus**, qui a permis d'améliorer la précision des deux ressources (Sagot et Fišer, 2012b), dont les résultats, qui seront bientôt distribués et le sont déjà en version bêta, seront le WOLF 1.0 et sloWNet 4.0. Nous avons donc développé pour celaune technique indépendante de la langue pour l'identification d'intrus dans les synsets, qui repose sur l'utilisation de corpus afin de désambiguïser au mieux les littéraux dans leurs contextes d'apparition (Sagot et Fišer, 2012b), en adaptant la notion de similarité distributionnelle entre un littéral et un wordnet évoquée brièvement ci-dessus. L'hypothèse sous-jacente est simple : les mots sémantiques, c'est-à-dire ici les couples *(littéral, synset)*, tendent à apparaître en corpus au voisinage d'autres mots sémantiques avec lesquels ils sont en relation sémantique, lesquelles relations sont celles qui sont modélisées par les liens entre synsets dans un wordnet. À ce jour, nous n'avons appliqué cette

25. C'est une double approximation : d'une part, les candidats déjà présents dans le WOLF ou dans sloWNet ne sont pas tous valides ; d'autre part, les candidats qui n'y sont pas encore ne sont pas tous erronés — en réalité, l'objectif ici est précisément d'identifier parmi eux ceux qui sont corrects.

26. Une évaluation manuelle de 400 des 177 980 candidats proposés pour WOLF, choisis aléatoirement, montre une précision moyenne de 52%, qui monte à 81% si l'on se restreint aux candidats faisant partie des 55 159 candidats retenus. À l'inverse, la précision parmi les candidats éliminés tombe à 40%. On note même que les trois derniers quartiles (si l'on classe les candidats en fonction du score que leur attribue le classifieur), qui correspondent à peu près aux candidats rejetés, correspondent à des précisions moyennes en chute libre : successivement 63%, 41% et 20%, ce qui montre la pertinence des scores associés aux candidats.

27. Appliqué aux 177 980 candidats *(littéral, synset)* français construits à partir de nos lexiques bilingues, et en utilisant le seuil à 0,1 mentionné ci-dessous, sont retenus 55 159 candidats dont 15 313 (28%) sont déjà dans la version précédente du WOLF. Autrement dit, la méthode propose 39 823 nouveaux candidats, augmentant de 43% le nombre de synsets non vides et de 65% le nombre de couples *(littéral, synset)*. Les chiffres concernant sloWNet sont encore plus éloquents, avec des pourcentages respectifs d'augmentation de 93% et de 244%. On se reportera au tableau 6.2 pour plus de détails.

approche qu'aux synsets nominaux, et ce en quatre étapes : (i) pour chaque nom attesté dans un corpus (monolingue) de taille importante et présent dans au moins un synset de la ressource à nettoyer, évaluation de la similarité sémantique entre le synset considéré et chaque occurrence du littéral, selon l'approche esquissée à l'étape précédente ; (ii) évaluation globale de chaque couple *(littéral, synset)* du wordnet à nettoyer à partir des similarités sémantiques calculées précédemment pour chaque occurrence du littéral ; (iii) extraction en tant que candidats intrus des couples *(littéral, synset)* dont le score global est en dessous d'un certain seuil ; (iv) tri manuel entre candidats intrus pour que soient retirées des wordnets les couples *(littéral, synset)* effectivement erronés. Des exemples de candidats intrus ainsi retenus, ainsi que leur évaluation manuelle, sont fournis pour le français et pour le slovène à la table 6.3 [28].

Certains travaux spécifiques à l'une ou l'autre des resssources ont également contribué à leur développement, que nous ne ferons qu'évoquer ici. Concernant le WOLF, entre la première et la deuxième étape, un travail spécifique a été réalisé sur les synsets adverbiaux qui tirait parti de relations de morphologie dérivationnelle et de la base de synonymes DicoSyn (Sagot *et al.*, 2009a,b). Le résultat de ce travail, qui a servi d'entrée à la deuxième étape, est la version 0.1.6.

Entre la seconde et la troisième étape, ou en parallèle à la seconde étape, le WOLF a été également étendu au moyen de techniques d'induction de sens non-supervisée à partir d'extraction de nominalisation reposant sur des corpus français analysés syntaxiquement (Gábor *et al.*, 2012), de corpus bilingues alignés (Apidianaki et Sagot, 2012, 2014) et d'un graphe de traduction fortement multilingue, YaMTG, extrait de ressources wiki (Hanoka et Sagot, 2012 ; Hanoka, 2015) [29]. De plus, une validation manuelle partielle des synsets verbaux BCS, et notamment de la quasi-totalité des BCS 1 verbaux, a été réalisée (correction et complétion) : 825 couples *(littéral, synset)* ont été ainsi ajoutés, 4 933 ont été supprimés et 5 204 ont été confirmés. De plus, des techniques d'identification d'erreurs plus simples que l'étape 3 ont été appliquées, afin de vérifier à partir du Le*fff* que les littéraux mono-tokens connus du Le*fff* l'étaient avec la partie du discours correspondant à leur synset. Une validation manuelle des littéraux ainsi identifiés comme suspects a permis de supprimer 4 155 couples *(littéral, synset)* incorrects.

De son côté, sloWNet a fait l'objet de plusieurs transformations entre la première et la seconde étape. Tout d'abord, parce qu'à ce moment-là il la situation juridique à propos du

28. Une interface de validation dédiée, nommée sloWCrowd, a été développée pour faciliter la validation manuelle de façon collaborative (Tavčar *et al.*, 2012). À ce jour, c'est pour sloWNet que sloWCrowd a été le plus utilisé, avec plusieurs dizaines de milliers de validations individuelles concernant plus de 7 000 candidats en juillet 2014. Pour le WOLF, nous avons obtenu plus de 6 000 validations individuelles permettant de confirmer ou d'éliminer 1 540 couples *(littéral, synset)* distincts.

29. YaMTG est librement disponible sous différentes versions sur la page de Valérie Hanoka, dont j'étais co-directeur de thèse, à l'URL suivante : `http://alpage.inria.fr/~hanoka/yamtg.html`.

dictionnaire bilingue anglais–slovène utilisé en complément des ressources wiki n'était pas clair, nous avons reproduit cette première étape sans le prendre en compte, donnant ainsi naissance à un wordnet plus petit que la version 2.0. Deuxièmement, nous n'avions pas réalisé que les wiki utilisaient parfois les marques d'accent sur les mots slovènes, ce qui est une information intéressante mais pas indiquée dans l'orthographe conventionnelle, d'où des doublets comme *kolo* et *koló*, que nous avons donc normalisés et fusionnés. Enfin, sloWNet ayant été utilisé pour une tâche d'annotation sémantique manuelle de corpus, il a fait l'objet d'une validation manuelle importante, correction d'erreurs et ajouter des nouveaux couples *(litéral, synset)* à la clef. Le résultat de ces modifications, qui a servi d'entrée à l'étape 2, est la version 2.2.

Enfin, on notera que c'est lors de la seconde étape, l'étape d'extension, que nous avons converti les identifiants de synsets depuis l'inventaire du PWN 2.0 à celui du PWN 3.0.

6.3 Évaluation des ressources

Comme pour toute ressource lexicale, évaluer des wordnets comme le WOLF ou sloWNet est une tâche délicate. Les trois paradigmes habituels, présentés à la section 5.3, restent valides : évaluation comparative par rapport à d'autres ressources comparables, évaluation manuelle de la précision voire de la couverture, et évaluation orientée-tâche. Cependant, à ce jour, le WOLF n'a pas encore fait l'objet d'une évaluation orientée-tâche, contrairement à sloWNet. Cette partie propose donc une évaluation semi-automatique de la qualité des ressources, qui s'appuie sur deux principes : (i) le FWN et le Slovene WordNet, développés manuellement, ne contiennent que des informations correctes (mais sont incomplètes), et les couples *(littéral, synset)* du WOLF et de sloWNet qui sont également présents respectivement dans le FWN et le Slovene Wordnet sont donc corrects ; (ii) un échantillonage aléatoire, partie du discours par partie du discours, des autres couples *(littéral, synset)*, permet d'estimer la qualité des autres couples, et donc *in fine* la qualité globale des ressources.

Le WOLF 1.0 et le sloWNet 4.0 n'étant pas encore disponibles, comme expliqué à la section précédente, les évaluations données dans cette partie concernent le résultat de l'étape d'extension, c'est-à-dire le WOLF 0.2 et sloWNet 3.0.

Le détail de la méthodologie et des résultats de ces évaluations peut être consulté dans (Sagot, 2017c) pour le WOLF et (Fišer et Sagot, 2015) pour sloWNet. En résumé, le WOLF 0.2 a une précision approximative de 86% (environ 65 700 couples *(littéral, synset)* sur 76 436). On peut comparer ce score avec ceux de WoNeF : la précision des 88 736 couples présents dans cette ressource a été évaluée avec soin par Pradet *et al.* (2014a)

(a) WOLF

	PWN 2.0	**WOLF 0.1.4**	**WOLF 0.2**	**WOLF 1.0b4**	**FWN**	**JAWS**	**WoNeF**
Total	115 424	32 351	46 449	56 479	22 121	34 367	53 442
BCS1	1 218	869	1 067	1 107	1 211	760	816
BCS2	3 471	1 665	2 519	2 900	3 022	1 729	2 097
BCS3	3 827	1 796	2 585	2 963	2 304	1 706	1 906
Hors BCS	106 908	27 492	40 278	49 509	15 584	30 172	48 623
N	79 689	28 187	36 933	42 427	17 381	34 367	37 355
V	13 508	1 546	4 105	5 870	4 740	0	3 845
Adj	18 563	1 422	4 282	6 691	0	0	10 238
Adv	3 664	667	1 125	1 487	0	0	2 002

(b) sloWNet

	PWN 2.0	**sloWNet 2.0**	**sloWNet 2.2**	**sloWNet 3.0**
Total	115 424	29 108	17 817	42 919
BCS1	1 218	714	1 203	1 208
BCS2	3 471	1 361	2 192	3 111
BCS3	3 827	1 611	1 232	2 698
Hors BCS	106 908	25 422	13 190	35 902
N	79 689	22 927	16 234	30 911
V	13 508	1 547	1 097	5 337
Adj	18 563	4 376	429	6 218
Adv	3 664	258	57	453

Tableau 6.2 – Données quantitatives (nombre de synsets non vides) sur les versions initiales du WOLF et de sloWNet (0.1.4 et 2.0 respectivement) et les versions obtenues après la phase d'extension (versions 0.2 et 3.0 respectivement), et la dernière version du WOLF (1.0b4, cf. plus bas). Sont également indiquées les données correspondantes pour le PWN (PWN), ainsi, pour le français, que les données pour le wordnet français développé dans le cadre du projet EuroWordNet (FWN) (Vossen, 1999), le wordnet nominal JAWS (Mouton et de Chalendar, 2010) et son successeur WoNeF (Pradet *et al.*, 2014a).

Candidats intrus trouvés dans le WOLF 0.2					Candidats intrus trouvés dans sloWNet 3.0				
littéral	synset	littéraux anglais du synset	score ($\times 10^3$)	éval	littéral	synset	littéraux anglais du synset	score ($\times 10^3$)	éval
abord	8307589	*meeting, group meeting*	0.013	OK	*aktiva*	5154517	*plus, asset*	0.002	OK
activité	14006945	*activeness, action, activity*	0.014	NO	*cilj*	5868477	*end*	0.004	OK
activité	5833022	*business*	0.011	OK	*dan*	15113229	*period, period of time, time period*	0.001	NO
adresse	35189	*achievement, accomplishment*	0.017	OK	*dan*	15157225	*day*	0.004	NO
agence	3015254	*chest, chest of drawers, chest, chest of drawers, bureau, dresser*	0.015	OK	*dan*	6210791	*light*	0.003	OK
besogne	6545137	*deed of conveyance, title deed*	0.012	OK	*dan*	6832572	*n, N*	0.004	OK
bout	8566028	*terminal, end*	0.019	NO	*datelj*	15159583	*date, day of the month*	0.000	OK
bureau	13945102	*office, power*	0.006	OK	*del*	5867413	*division, part, section*	0.003	NO
cadre	10069645	*executive director executive*	0.017	OK	*del*	13809207	*constituent, component, component part, part, portion*	0.003	NO
cadre	10014939	*managing director, manager, director*	0.014	OK	*delež*	5256358	*part, parting*	0.004	OK

Tableau 6.3 – Exemple d'intrus identifiés dans le WOLF 0.2 et dans sloWNet 3.0, accompagnés d'une évaluation manuelle. Il s'agit des 10 premiers candidats des deux ensembles de candidats (un par langue) extraits par sélection aléatoire à des fins d'évaluation manuelle de notre approche. La tâche consistant à identifier des erreurs, « OK » indique une erreur, correctement identifiée, alors que « NO » indique un couple correct et dont l'identification comme erreur est donc invalide.

comme étant de 68,9% [30]. Concernant sloWNet, on obient des résultats similaires : environ 70 690 couples corrects sur 82 721 soit une précision de l'ordre de 85%.

On pourra se reporter aux différentes publications citées au cours du chapitre pour des évaluations complémentaires, notamment de la qualité des intrus proposés par l'étape 3. De nombreuses évaluations sont notamment disponibles concernant sloWNet dans (Fišer et Sagot, 2015), y compris une évaluation orientée-tâche dans un contexte de traduction automatique, une tentative d'évaluation du rappel de la ressource (chiffre bien plus difficile à estimer que la précision) et une évaluation comparative avec les wordnets multilingues UWN (de Melo et Weikum, 2009) et BabelNet 2.0 (Navigli et Ponzetto, 2010, 2012), qui montrent de façon manifeste que sloWNet est actuellement et de loin le wordnet le plus utilisable pour le slovène [31].

6.4 Travaux en cours et perspectives

Le travail de validation manuelle des candidats intrus se poursuit. C'est donc à court terme que devraient être publiées les versions 1.0 de WOLF et 4.0 de sloWNet, qui en intègreront les résultats. C'est d'autant plus souhaitable que les taux d'erreurs constatés sur les wordnets dans lesquels nous avons cherché à détecter des intrus sont bien plus élevés dans les synsets de base que dans les synsets les plus spécifiques, comme le montre le précision de 92% obtenue sur les couples *(littéral, synset)* obtenus par le WOLF sur les synsets non couverts par le FWN, qui sont donc les moins fréquents, ou comme indiqué pour le slovène par Fišer et Sagot (2015). Or ces synsets de base auront vraisemblablement tendance à jouer un plus grand rôle dans toute intégration de ces ressources au sein d'outils tels que des systèmes de traduction automatique, d'extraction d'information ou d'annotation sémantique profonde.

WoNeF et WOLF étant tous deux alignés sur le PWN, leur fusion peut être effectuée directement. Ce travail a déjà été réalisé à partir de la version 1.0b2 du WOLF, et le résultat a été évalué de façon préliminaire. Nous prévoyons de poursuivre dans cette voie.

Enfin, au-delà de leur peuplement, ces ressources devront être enrichies, notamment par l'ajout de définitions et d'exemples d'usage, ou encore par l'ajout de traits sémantiques génériques (ainsi *animé* ou *abstrait* pour les synsets nominaux, ou encore *verbe de*

30. Ces auteurs ont également proposé un score de rappel, qui est de 68,9%. Une version orientée-précision de WoNeF est également distribuée. Elle atteint 91,5% de précision, mais ne contient que 15 625 couples (couverture estimée à 51,5%).

31. Étant données les trois ressources que sont sloWNet 3.0, UWN et BabelNet 2.0, nous avons évalué à la main 350 couples *(littéral, synset)* tels que chaque combinaison de présence ou absence de chacune des trois ressources était couverte par 50 couples. Ainsi, il y avait par exemple 50 couples présents dans sloWNet 3.0 et UWN mais pas dans BabelNet 2.0, 50 couples présents un uniquement dans BabelNet 2.0, et ainsi de suite. On se reportera à (Fišer et Sagot, 2015) pour les résultats détaillés, mais en résumé, sloWNet 3.0 contient 82 721 couples dont 88% sont corrects (pour cette évaluation), BabelNet 2.0 en contient plus, 131 964, mais avec une précision de seulement 73%, et UWN a une précision comparable à sloWNet, 87%, mais ne contient que 9 924 couples.

mouvement ou *verbe d'état* pour les verbes). Par ailleurs, pour le français, le couplage du WOLF avec d'autres ressources, parmi lesquelles le Le*fff* et le lexique partiel à la FrameNet développé pour le français dans le cadre du projet ASFALDA (Candito *et al.*, 2014) [32], donnera accès à une vision plus transversale du lexique de cette langue. Enfin, le développement de divers lexiques dérivationnels, au niveau morphologique voire au niveau morphosémantique, devrait permettre l'ajout de liens dérivationnels entre synsets voire entre couples *(littéral, synset)*, ouvrant ainsi des perspectives intéressantes tant sur le plan linguistique que sur le plan du traitement automatique des langues.

32. https://sites.google.com/site/anrasfalda/

Chapitre 7

Analyse de surface et informations lexicales

Sommaire

7.1 Introduction

7.1.1 Problématique

De nombreux travaux en traitement automatique des langues, y compris certains des nôtres qui seront présentés aux chapitres suivants, partent d'une hypothèse très forte, celle selon laquelle les données que l'on doit traiter sont déjà segmentées en unités de traitement, les « phrases », elles-mêmes segmentées en unités élémentaires pertinentes. Or, comme nous l'avons discuté au chapitre 1, ces unités élémentaires sont, pour l'étiquetage morphosyntaxique comme pour l'analyse syntaxique, ce que nous avons appelé des *mots syntaxiques*. Ces unités sont les récipiendaires légitimes des informations de partie du discours et constituent les nœuds terminaux des structures syntaxiques telles que les arbres de constituance ou de dépendance syntaxique.

Naturellement, les outils de TAL ne peuvent pas *a priori* se contenter de faire une hypothèse aussi forte que celle-ci. Dans les situations réelles, ces outils ont à traiter des données textuelles brutes non segmentées, parfois même bruitées (notamment orthographiquement et typographiquement). Comme nous l'avons évoqué au chapitre 1, les mots syntaxiques ne sont pas systématiquement en relation bijective avec les *mots typographiques*, ou *tokens*,[1] seules unités qu'il est possible d'identifier automatiquement dans des textes bruts sans risque d'erreur (cf. section 1.3.7). Il est donc indispensable, pour traiter des corpus réels, de disposer de moyens de segmenter ceux-ci en « phrases » puis d'extraire pour chacune d'entre elles la séquence de mots syntaxiques dont elle est constituée, ou du moins une approximation éventuellement non déterministe de cette séquence, à partir des seules unités facilement identifiables, les tokens.

La correspondance entre tokens et mots syntaxiques n'est pas toujours bijective. En revanche, il s'avère que la correspondance entre *formes* (ou *mots morphologiques*) et mots syntaxiques est bien plus systématique, du moins dans les langues que nous allons traiter (cf. chapitre 1). Nous nous appuierons donc dorénavant sur l'approximation consistant à identifier le problème de l'identification des mots syntaxiques et celui de l'identification des formes. L'identification des « phrases » et des formes à partir d'un corpus bruts, en passant par une étape préliminaire de tokenisation, est l'objet de ce chapitre.

Le tableau 7.1 montre sur des exemples du français la non-correspondance entre tokens et formes, illustrant ainsi le cas d'une langue faisant usage de séparateurs typographiques. Ce tableau illustre certains des phénomènes sous-jacents à cette non-correspondance, en indiquant à chaque fois la forme ou le treillis de formes correspondant à un ou plusieurs tokens. Elle illustre la diversité de ces phénomènes, rarement considérés comme relevant d'une même problématique générale. On notera qu'une non-correspondance entre tokens et formes n'implique pas nécessairement une différence entre nombre de tokens et nombre

1. En anglais, on trouve également le terme de *source token*.

Tokens	Treillis de formes	Commentaire
	Aucune altération	
a priori	*a_priori*	forme composée
aujourd'hui	*aujourd'hui*	forme simple
l' idée	*l' idée*	deux formes simples
pomme de terre	*(pomme de terre | pomme_de_terre)*	ambiguïté entre formes simples et forme composée
duquel	*de lequel*	amalgame
à l' instar du	*à_l'_instar_de le*	interaction entre composé et amalgame
grâce au	*(grâce à | grâce_à) le*	idem, avec ambiguïté
pré-commande -t-il	*pré-_ commande -t-il*	préfixe productif
www.inria.fr	*_URL*	forme spéciale mono-token
12 345 , 6	*_NUMBER*	forme spéciale multi-tokens
Vyacheslav Ivanov	*_PERSON_m*	idem (nom propre)
; -)	*_SMILEY*	émoticône
	Altérations ergographiques	
c' est l' **ete**	*c' est l'* ***été***	diacritique manquant
math	***mathématiques***	troncation
nimp	***n'importe quoi***	troncation et contraction
je **sé**	*je* ***sais***	réduction phonétique
c a dire	***c'est-à-dire***	réduction phonétique et surdécoupage
c t	***c' était***	réd. phonétique, contraction et surdécoupage
je suis **oqp**	*je suis* ***occupé***	réduction symbolique
2m1	***demain***	réduction symbolique
bjr	***bonjour***	squelette consonnantique
son normaux	***sont*** *normaux*	autre erreur orthographique
sqlut	***salut***	faute de frappe
A1203	***Al_2O3***	erreurs d'OCR
	Altérations expressives	
Joli !!!!!!	*Joli !*	étirement de ponctuation
superrrrrrrr	***super***	étirement graphémique
in-cro-ya-ble	***incroyable***	décomposition
m*e**	***merde***	autocensure
très joli	***très*** *joli*	formatage typographique

Tableau 7.1 – Quelques exemples en français du passage des tokens aux formes. Sont appliquées les conventions de tokenisation utilisées par défaut par SxPipe pour le français (cf. section 7.1.2 : dans la première colonne, les espaces sont les frontières entre tokens). Pour les différents types d'altérations, que l'on trouve notamment sur le web, cf. section 7.3.3 et les références citées. Les tokens résultant d'une altération sont indiqués en gras, ainsi que les formes correspondantes.

de formes. Ces catégories, qui, bien sûr, ne sont pas mutuellement exclusives, sont les suivantes :

— les unités multi-tokens, y compris les composés [2] et les entités nommées multi-tokens ;

— les amalgames (par exemple en français `aux` vs. *à les*),

— les entités nommées au sens large (mono- ou multi-tokens),

— les altérations, qu'elles soient volontaires (réductions volontaires, altérations expressives…) ou non (erreurs d'orthographe, fautes de frappe…).

L'objectif de ce chapitre est donc d'explorer les différentes approches que l'on peut mettre en œuvre pour contourner, approcher ou, au moins en partie, traiter l'écart qu'il y a entre données tokenisées et données analysées en formes. Nous évoquerons ainsi, à des degrés divers, les tâches suivantes :

— l'identification des composés, que nous verrons principalement sous l'angle de l'identification des frontières entre formes dans des données en mandarin, à l'aide d'une approche non supervisée (section 7.4),

— le traitement des entités nommées (c'est ici l'exception : nous ne revenons pas dans ce documents sur nos travaux concernant les entités nommées, lesquels sont évoqués ci-dessous, avec références, dans la note 21),

— la correction ou la normalisation orthographique — la différence sera définie plus bas —, notamment dans le contexte des données issues du web (section 7.3),

Comme discuté ci-dessus, ces tâches nécessitent au prélable une segmentation en « phrases » et en tokens (section 7.2).

7.1.2 SxPipe

Une partie importante de notre travail sur ces sujets a été mise en œuvre au sein de SxPipe, chaîne modulaire et multilingue de traitements de surface (Sagot et Boullier, 2005a,b, 2008) développée au départ comme préliminaire pour l'analyse syntaxique symbolique par les analyseurs FRMG et SxLfg (cf. chapitre 9), initialement en vue de la campagne EASy d'évaluation des analyseurs syntaxiques (Boullier *et al.*, 2005a ; Paroubek *et al.*, 2006) [3].

2. Rappelons que nous avons montré au chapitre 1 comment définir la notion de token dans les systèmes d'écriture sans séparateur typographique. Par exemple, dans les systèmes faisant usage des sinogrammes, chaque sinogramme est un token. Ainsi, le terme de *composé* étant défini comme le cas où plusieurs tokens participent à la transcription de la même forme, tout « mot » mandarin transcrit par plusieurs sinogrammes est formellement un composé.

3. Toutes les versions de SxPipe sont accessibles librement sous licence Cecill-C (compatible LGPL), au sein du projet `lingwb` de la GForge Inria (cf. `http://gforge.inria.fr/projects/lingwb/`).

De nombreuses chaînes de traitement réalisant les mêmes types de tâches que SxPipe existaient naturellement déjà au moment du lancement du développement de SxPipe, parmi lesquelles Unitex (Paumier, 2003), GATE (Cunningham *et al.*, 2002) et LinguaStream (Bilhaut et Widlöcher, 2006). Nous avions néanmoins constaté que les outils disponibles n'étaient pas toujours satisfaisants pour une application donnée, notamment parce qu'ils ne disposaient pas d'un module de correction orthographique, parce qu'ils n'étaient adaptés qu'à tel ou tel type de corpus, ou parce qu'ils ne savaient pas gérer de façon satisfaisante les ambiguïtés qui peuvent apparaître à chaque étape (Sagot et Boullier, 2008).

C'est pour répondre à ces insuffisances que nous avons développé SxPipe. Ce système permet de transformer un texte brut en un treillis de formes, entrée potentiellement pertinente pour différents types d'outils, et par exemple pour un analyseur syntaxique, pour un outil de normalisation textuelle (par re-transformation en texte brut), ou pour un système d'extraction d'informations *via* des entités nommées. Dès la première version (Sagot et Boullier, 2005a,b), SxPipe permettait le traitement de corpus variés en langue française, à l'image des différents types de sous-corpus utilisés au cours de la campagne EASy, grâce notamment à deux types de modules : (i) un ensemble de deux modules de correction orthographique et d'identification des formes composées, dont le second était non déterministe, et (ii) des modules de reconnaissance de divers types d'entités nommées et d'autres types de motifs. Enfin, SxPipe permet la préservation des ambiguïtés dès lors qu'un module ne dispose pas de suffisamment d'informations pour choisir une seule interprétation des données.

La deuxième version de SxPipe, présentée dans (Sagot et Boullier, 2008), présentait plusieurs évolutions : (i) le passage au multilingue, notamment par le traitement du polonais (Sagot, 2007) mais aussi de l'anglais, de l'espagnol et du slovaque ; (ii) un module unifié de correction orthographique et de reconnaissance de formes composées, qui peut produire des sorties ambiguës : la correction orthographique, lieu d'incertitudes aussi important que la reconnaissance des mots composés, peut dès lors se faire de façon complètement non déterministe ; (iii) une architecture originale pour la reconnaissance de motifs décrits à l'aide de grammaires non contextuelles (CFG) ; (iv) de nombreuses améliorations apportées aux divers composants de la chaîne en termes de couverture, de précision et de vitesse d'exécution. Pour plus de détails, on pourra se référer à (Sagot et Boullier, 2008).

Ce développement s'est poursuivi depuis, et l'architecture mise en place a hébergé plusieurs travaux, dont nous allons évoquer certains ce chapitre.

SxPipe est un système modulaire : une chaîne de traitement est définie par un fichier de configuration spécifiant les modules à utiliser et les options à donner à chaque module. Les options données sur la ligne de commande à l'exécution sont transmises à ceux des modules pour lesquels la configuration le prévoit. SxPipe est fourni avec une

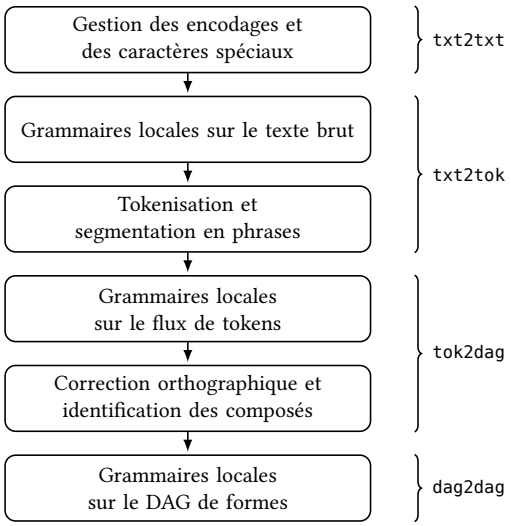


Figure 7.1 – Architecture générale de la chaîne de traitement SxPipe. Les sous-dossiers du paquatage SxPipe sont indiqués à droite du schéma pour information.

configuration standard, mais également avec quelques configurations adaptées à certains types de corpus, certains usages ou certains ensembles de langues. Cependant, les modules ne peuvent être agencés dans n'importe quel ordre, chacun d'entre eux appartenant structurellement à l'une des 5 étapes indiquées à la figure 7.1, et étant parfois prévus pour être placés avant ou après certains autres modules.

7.2 Tokenisation, segmentation en phrases et détection des entités typographiques

7.2.1 Tokens

Nous avons défini à la section 1.3.7 la notion de mot typographique, ou token. Pour les systèmes d'écriture disposant d'un séparateur typographique (souvent l'espace) un token est généralement défini comme étant soit une séquence contiguë de caractères délimitée de part et d'autre par un séparateur ou un caractère de ponctuation, soit un caractère de ponctuation.

Une telle définition suppose de définir ce qu'est un symbole de ponctuation. Par ailleurs, pour les systèmes d'écriture de disposant pas de sépérateur typographique, cette définition n'est pas opératoire. Pour de tels systèmes d'écriture, nous avons vu à la section 1.3.7 que l'on pouvait souvent identifier des unités typographiques élémentaires qui sont généralement ou plus petites (caractères) ou plus grandes (séquences de mots

délimités par des ponctuations) qu'un token standard tel qu'obtenu avec un système d'écriture disposant d'un séparateur.

Une conséquence de cette différence est que les techniques qu'il convient de mettre en œuvre pour identifier les formes sous-jacentes à la séquence de tokens constituant un énoncé ne sont pas forcément les mêmes dans les deux cas. Dans le cas des systèmes d'écriture avec séparateur typographique, la majorité des tokens correspondent finalement plus ou moins à des formes simples, et la tâche consiste donc notamment à identifier les formes composées et les amalgames. En l'absence de séparateur typographique, le choix le plus naturel est de considérer chaque caractère comme un token, avec pour conséquence que la majorité des formes sont constituées de plusieurs tokens. Un point est néanmoins commun à tous les cas : certaines entités nommées très conventionnelles, dont la structure interne est définie au niveau des caractères, gagnent à être détectées avant la tokenisation. C'est le cas notamment des dates et des horaires (sauf ceux écrites en toutes lettres), des nombres et mesures (même remarque), des URLs et adresses e-mail, des émoticônes (ou *smileys*), et plus généralement de toutes les entités que nous qualifierons d'*entités typographiques*.

7.2.2 Phrases

La quasi-totalité des systèmes de traitement automatique des langues requièrent un découpage préalable en unités de traitement que nous appelons par commodité des « phrases ». Si par exemple on souhaite réaliser ensuite une analyse syntaxique reposant sur une grammaire, il convient de chercher à faire correspondre les phrases obtenues par l'étape de découpage avec les phrases appartenant au langage défini par la grammaire : un terme plus précis serait donc celui de *phrase grammaticale*[4]. Au niveau typographique, la plupart des systèmes d'écriture disposent de marqueurs de fin de phrase, ou ponctuations finales[5], dont certains donnent en plus une information sur la nature de la phrase qu'ils terminent (interrogative[6], exclamative, etc.) ou indiquent une coupure plus

4. Au plan linguistique, cette définition de la phrase (grammaticale) correspond approximativement à celle d'*unité rectionnelle* chez Berrendonner (2002) ou Benzitoun *et al.* (2010).

5. La ponctuation finale non marquée, dans la majorité des systèmes d'écriture actuels, est le point (systèmes reposant sur les alphabets latin, grec et cyrillique, systèmes reposant sur les alphasyllabaires arabe et hébreu, hangeul utilisé pour noter le coréen). Dans les systèmes utilisant les sinogrammes (chinois, japonais), le caractère de fin de phrase est « 。 ». Certains systèmes d'écriture utilisent d'autres symboles, comme par exemple le *verjaket* (« ։ ») dans le système arménien, le *daṇḍa* (« । ») en devanagari (utilisé pour noter de nombreuses langues de la péninsule indienne dont le hindi, le népali et le sanskrit), le *'arat nettib* (lit. quatre-points, « ፨ ») dans l'alphasyllabaire guèze, utilisé notamment pour transcrire l'amharique et le tigrigna, ou le *khăn* (« ។ ») dans le système khmer.

6. La fin d'une phrase interrogative est marquée par le point d'interrogation dans la plupart des systèmes d'écriture. Ce point d'interrogation est inversé dans le système arabe et ses dérivés, où le sens d'écriture est de droite à gauche : « ؟ ».

macroscopique de l'énoncé [7]. Certains systèmes d'écriture ont également des marqueurs de début de phrases, au moins dans certains cas [8].

Mais certains de ces marqueurs typographiques sont également utilisés à d'autres fins, et notamment le point dans les langues occidentales dont le français, induisant un décalage entre le balisage typographique des phrases et la notion de phrase grammaticale. Cette difficulté est rencontrée au moins dans les deux situations suivantes :

— Lorsqu'un marqueur typographique de changement de phrase peut également avoir d'autres rôles (ainsi le point dans les acronymes) ;

— Lorsque l'on utilise un marqueur typographique de changement de phrase pour créer un effet stylistique (et le cas échéant prosodique). Ce sont des cas que nous avons étudiés avant tout sur le français (Danlos et Sagot, 2010) et nommés « ponctuations fortes abusives » [9]. Un des exemples cités dans (Danlos et Sagot, 2010), tiré du corpus de l'Est Républicain (Gaiffe et Nehbi, 2009 ; Seddah *et al.*, 2012a), est le suivant : *Des défis qui engagent son avenir. Et son devenir.* Autre type d'exemple, en anglais cette fois : *BEST. MOVIE. EVER.* (Seddah *et al.*, 2012c).

Il existe une autre difficulté plus fondamentale : la notion de phrase telle que nous l'avons formulée ci-dessus, c'est-à-dire la notion de phrase grammaticale, pose parfois problème : sa définition même implique qu'il n'y ait pas de relation syntaxique dépassant le cadre de la phrase, et donc que toute relation inter-phrastique ne saurait être syntaxique. Il y a au moins deux cas où un tel postulat est approximatif :

— Les listes, d'une part parce qu'elle induisent une structuration qui peut s'entrecroiser avec la structure syntaxique, mais surtout parce qu'elles peuvent remplir un argument syntaxique tout en contenant des frontières de phrases (le deuxième élément de la liste précédente en est un exemple, qui illustre également le fait qu'un élément de liste peut commencer par une majuscule) ;

— Les citations sous forme de discours rapporté, les guillemets délimitant des segments qui peuvent contenir des fontières de phrase, y compris dans des exemples où la citation est en relation avec un verbe mis en incise, à l'intérieur au voisinage immédiat de la citation (Sagot *et al.*, 2010 ; Sagot et Danlos, 2010 ; Danlos *et al.*, 2010). Un exemple attesté [10] fourni dans (Danlos *et al.*, 2010) illustre ce phénomène : *« J'ai trouvé (sa) colère suspecte et préparée. Je suis outré par ce que ce qui a été dit, c'est des mensonges », a-t-il ajouté.* Nous avons néanmoins proposé dans (Danlos *et al.*, 2010)

7. Ainsi le double daṇḍa (« ॥ ») du devanagari, le *bâriyaôsan* (« ᩞ ») du système khmer, le séparateur de paragraphes du système guèze (« ፠ »), et d'autres.

8. Ainsi, en espagnol, les points d'interrogation et d'exclamation inversés (« ¿ » et « ¡ ») qui sont placés en début de phrase, en plus des points d'interrogation et d'exclamation usuels placés en fin de phrase.

9. Dans les études sur l'oral, où les signes de ponctuation cèdent la place aux marqueurs prosodiques, ce phénomène correspond *grosso modo* à la notion d'épexégèse, terme introduit par Bailly (1944).

10. AFP, dépêche TX-SGE-UXM34 du 3 mai 2007

une analyse et une modélisation formelle au niveau discursif (inter-phrastique) et non syntaxique (intra-phrastique) du lien qu'entretient une citation avec le verbe de citation qui lui correspond, et ce pour des raisons avant tout linguistiques.

7.2.3 Tokenisation et segmentation en phrases dans SxPipe

La version actuelle de SxPipe inclut un tokeniseur multilingue qui effectue de façon jointe le découpage en tokens et en phrases pour un grand nombre de langues et de systèmes d'écriture, avec ou sans séparateur typographique. Ce tokeniseur explicite un certain nombre de conventions contextuelles qui, pour certaines, dépendent de la langue [11], y compris pour normaliser certaines pratiques typographiques spécifiques [12]. Seules certaines langues bénéficient de conventions raffinées par rapport aux conventions génériques. Réaliser conjointement la tokenisation et la segmentation en phrases permet d'identifier au mieux les cas où les marqueurs pouvant être utilisés pour délimiter des phrases le sont à d'autres fins.

Comme indiqué ci-dessus, les étapes de tokenisation et de segmentation en phrases gagnent à n'être exécutées qu'après que certains types d'entités, les entités que nous avons qualifiées de typographiques, ont été reconues [13]. Il s'agit d'entités qui font usage dans leur structure interne des signes de ponctuation ou du séparateur typographique d'une façon qui ne correspond pas à leur rôle dans le système d'écriture en général, y compris des signes marquant la fin des phrases. Par exemple, pour une URL, il est plus simple de procéder à sa détection avant la tokenisation et la segmentation en phrases, de façon à ne pas considerer les points et les barres obliques qu'elle comporte comme des ponctuations, mais bien l'ensemble de l'URL comme un token unique.

Ainsi le module de tokenisation et segmentation en phrases de SxPipe est précédé, dans la configuration standard, d'un certain nombre de modules dont certains sont partiellement dépendants de la langue ou du système d'écriture :

— Quatre modules préliminaires de normalisation de l'encodage [14] ;

— Reconnaissance robuste des adresses e-mail, des URL, des dates (sauf celles écrites en toutes lettres), des indications horaires (sauf celles écrites en toutes lettres), des numéros de téléphone, des adresses postales [15] ;

11. En particulier, différentes conventions peuvent être activées concernant les tirets (quelle que soit la langue) et les apostrophes (pour le français seulement car les autres langues, bien que traitées de façon différentes d'une langue à l'autre, ne nécessitent pas de pouvoir choisir entre plusieurs conventions).

12. Par exemple, en italien, une apostrophe suivant une voyelle à la fin d'un token est presque toujours une manière d'accentuer cette voyelle : *piu'* est à lire *più* 'plus, davantage'.

13. Ceci implique des conventions définissant la notion de token qui soient en cohérence avec de tels choix.

14. Mise en UTF-8 valide, normalisation des caractères arabes pour éliminer les variantes contextuelles, échappement des caractères utilisés comme méta-caractères par SxPipe, restauration des caractères encodés sous forme d'entités XML.

15. Certaines des grammaires locales permettant la reconnaissance de ces entités nommées ne couvrent de façon complète que certaines langues.

— Reconnaissance de différents mécanismes typographiques spécifiques (ponctuations spéciales multi-tokens…) et de certaines spécificités des corpus d'oral transcrit [16] ;

— Reconnaissance des préfixes numériques [17] puis des nombres (y compris les fractions, etc.), des tokens alphanumériques (exemple : *A380*) et des en-têtes d'éléments de listes (exemples : -, *1)*, *A.b.*, ②, •, et d'autres) ;

— Reconnaissance des émoticônes et gestion des indications de formatage parfois utilisées dans le texte brut (exemples : *_mot_*, **mot**).

D'anciennes versions de certains de ces modules ont été évaluées en précision et en rappel sur des données du français (Sagot et Boullier, 2008).

D'autres modules ont été développés récemment, notamment dans le contexte du traitement des sorties de systèmes d'OCR, et évalués également en précision et rappel (Gábor et Sagot, 2014 ; Sagot et Gábor, 2014). Ces modules, que l'on peut insérer après la reconnaissance des adresses postales, sont dédiés successivement à la reconnaissance robuste des indications monétaires, des dimensions, des références juridiques et des formules chimiques [18], [19].

7.3 Correction et normalisation : le cas des systèmes d'écriture à séparateur typographique

Dans la suite de ce chapitre, nous supposerons donc effectuées les tâches de tokenisation et de segmentation en « phrases ».

Dans le cas des textes écrits dans un système d'écriture faisant usage de séparateurs typographiques, l'identification des formes relève de trois tâches distinctes mais fortement entremêlées : l'identification des composés (plusieurs tokens dans une même forme), des amalgames (plusieurs formes dans un même token, y compris des affixes productifs si l'on souhaite les traiter comme des unités syntaxiques autonomes) et des entités nommées de toute nature (qui peuvent relever des deux cas précédents tout en ayant leurs spécificités propres, dont la moindre n'est pas leur productivité) [20]. Nous avons déjà évoqué le cas

16. Par exemple les marques d'hésitation, répétitions, certaines disfluences simples, etc.

17. Pour des cas comme l'adjectif *5-foliés* dans les langues utilisant l'alphabet latin.

18. Ces modules sont en partie indépendants de la langue et en partie adaptés aux spécificités de certaines langues. Pour l'instant, ces adaptations ne concernent que certaines langues occidentales s'écrivant avec l'alphabet latin, voire uniquement au français dans le cas des références juridiques.

19. Rappelons que la modularité de SxPipe permet à chaque utilisateur de choisir les modules qu'il souhaite appliquer, dans quel ordre et avec quelles options, dont certaines peuvent être modifiées ou précisées à l'exécution. Ainsi, chaque type d'utilisation ou de corpus peut donner lieu à des paramétrages différents. De plus, les modules décrits dans cette section, y compris le module de tokenisation et segmentation en phrases, reconnaissent une option permettant de respecter une tokenisation pré-existante donnée en entrée.

20. Un cas particulièrement délicat est celui où un amalgame recouvre, entre autres, une partie d'un nom propre. C'est par exemple le cas dans une phrase comme « *Nous venons du Mans*, où les tokens `du Mans`

des entités typographiques comme les URL ou les nombres, deux catégories présentes dans le tableau 7.1. Nous ne discuterons pas ici du cas des noms propres, sur lequel nous avons toutefois travaillé, comme déjà mentionné à la section 3.2.2 [21]. Entre tokenisation et identification des noms propres, SxPipe reconnaît certains autres types de motifs, et notamment les séquences en langue étrangère (titres d'ouvrages ou de films, notamment), comme décrit plus en détails dans (Sagot et Boullier, 2008).

Restent donc les composés et les amalgames, deux cas qui peuvent s'entremêler comme illustré à la figure 7.1 par l'exemple `à l' instar du` vs. *à_l'_instar_de le*. On pourrait penser qu'il suffit de s'appuyer sur un lexique à large couverture comme le Le*fff* pour résoudre ce problème, moyennant quelques subtilités algorithmiques et de programmation pour que les temps de traitement restent bas (Sagot et Boullier, 2008). Toutefois, il est une difficulté supplémentaire, celle des altérations, qu'elles soient involontaires (fautes de frappe, d'orthographe, d'OCR ou autre) ou volontaires. Il faut en effet savoir distinguer un token qui est le résultat de l'accumulation de formes correctes (par amalgame, affixation et/ou apposition) d'un token qui est le résultat d'une ou de plusieurs altérations. La situation se complique lorsque l'on veut pouvoir gérer correctement les tokens qui sont à la fois le résultat de l'accumulation de plusieurs formes et d'une ou de plusieurs erreurs (par exemple, *done-moi* voire *redone-moi*, si REDONNER n'était pas connu du Le*fff*). Tout ceci est encore plus délicat lorsque l'on souhaite reconnaître les possibles formes composées (*pome de terre*), voire les espaces (frontières de tokens) insérés par erreur dans le texte brut.

On pourra se reporter à la section A.8 pour un rapide survol des travaux en correction orthographique. Dans la suite de cette section, nous évoquons tout d'abord la façon dont fonctionne l'outil TEXT2DAG intégré à SxPipe qui procède de façon jointe à l'identification des formes à partir des tokens et à leur correction orthographique grâce à des informations lexicales, ces deux tâches pouvant être réalisées de façon ambiguë. Nous nous pencherons ensuite sur une approche permettant l'extraction de règles de correction pondérée à partir de corpus, à partir d'une approche par analogie, et l'utilisation de ces règles au sein d'un système de correction robuste. Nous décrirons enfin brièvement deux approches complémentaires aux précédentes, qui se limitent à la correction des tokens, en prenant en compte leur contexte, mais sans aller jusqu'à l'identification des formes. Comme nous

correspondent successivement à la forme *de* et à l'entité nommée *Le Mans*, que SxPipe a vocation à considérer comme une forme spéciale _LOCATION. Ce cas précis n'est actuellement pas traité correctement par SxPipe.

21. Rappelons ici que nos travaux sur la reconnaissance et le liage d'entités nommées ont été effectués en grande partie dans le cadre des projets SCRIBO (projet du pôle de compétitivité System@tic) et EDyLex (projet ANR dont j'étais le porteur) ainsi que dans celui de la thèse de Rosa Stern (Stern, 2015), thèse CIFRE en partenariat avec l'Agence France-Presse dont j'étais l'encadrant principal et dont Denis Teyssou était le responsable scientifique en entreprise. Ils ont donné lieu à une série de publications (Villemonte de La Clergerie *et al.*, 2009b ; Stern et Sagot, 2010a,b ; Béchet *et al.*, 2011 ; Stern *et al.*, 2012 ; Sagot et Stern, 2012 ; Stern et Sagot, 2012 ; Sagot *et al.*, 2012, 2013).

le verrons, elles ont toutefois des avantages par rapport aux approches reposant sur la notion de forme et sont adaptées à des types de corpus différents.

7.3.1 Correction lexicale non déterministe par règles opérant au niveau des caractères et développées manuellement

La façon la plus globale de traiter les problèmes d'identification des formes à partir d'un flux de tokens tout en corrigeant les différents types d'altérations est procéder de façon jointe, en préservant les ambiguïtés dès lors que l'on ne dispose pas des informations permettant de les lever. C'est le rôle du module TEXT2DAG intégré à SxPipe, module décrit notamment dans (Sagot, 2006) et (Sagot et Boullier, 2008). TEXT2DAG repose sur un correcteur lexical encapsulé dans des heuristiques pour identifier notamment les clitiques et les appositions (cf. *scénario-catastrophe*, *donne-moi* ou *concept-clé*). Le correcteur lexical proprement dit fonctionne grâce à un système de règles de substitution pondérées permettant de remplacer des séquences d'un ou plusieurs caractères par une autre séquence d'un ou plusieurs caractères. Ces règles sont créées et pondérées manuellement et rassemblées en ensembles de règles. Chacun de ces ensembles, que l'on peut activer ou non, couvre un type cohérent de fautes (diacritiques manquants, diacritiques incorrects, caractères doublés ou non doublés par erreur, remplacement d'une séquence de caractères par une autre à la phonétique identique, fautes de proximité clavier, règles génériques d'insertion, délétion ou substitution…). Certaines de ces règles ou certains de ces jeux de règles dépendent de la langue. Sur option, l'insertion d'une espace ou la suppression d'une espace entre deux tokens peuvent être tentées. Le coût d'une correction est calculé à partir du coût des règles appliquées pour l'obtenir [22], et l'on peut indiquer un coût maximum autorisé et/ou un nombre maximum de correction à proposer pour un même token.

Grâce à une implémentation efficace faisant usage des structures de données du système SYNTAX (Boullier et Deschamp, 1988–2007), d'une représentation originale et efficace du lexique utilisé et d'un parcours parallèle de l'automate de l'input à corriger et de celui du lexique, l'exécution de TEXT2DAG est rapide. Par ailleurs, TEXT2DAG est facilement adaptable à une autre langue, pour peu que l'on dispose d'un lexique orthographique pour la langue donnée, les règles génériques indépendantes de la langue donnant déjà généralement des résultats raisonnables. Il a ainsi été utilisé pour d'autres langues que le français, et notamment le polonais (Sagot, 2007) et le persan (Sagot et Walther, 2010a ; Sagot *et al.*, 2011b,c), grâce aux lexiques Alexina PolLex et PerLex.

22. Le calcul du coût est plus complexe qu'une simple addition. On se référera à (Sagot et Boullier, 2008) pour plus de détails.

7.3.2 Correction et normalisation lexicales non déterministes par règles opérant au niveau des caractères et extraites automatiquement par analogie [23]

Le fait que les règles soient développées manuellement et, plus encore, que leurs poids soient également attribués manuellement sont d'évidentes limitations de TEXT2DAG. Nous avons donc étudié d'autres moyens de développer et de pondérer de telles règles. Nous avons pour cela mis en œuvre une technique d'apprentissage supervisé de règles pondérées qui repose sur l'analogie. Cette technique est une variante supervisée de la technique non supervisée décrite brièvement à la section 3.3.2 pour l'acquisition de liens dérivationnels au sein d'un lexique flexionnel et adaptée à la section 3.2.3 pour l'analyse morphologique de néologismes dérivationnels.

Cette technique, comme le correcteur lexical encapsulé dans TEXT2DAG, opère au niveau des tokens et n'est appliquée que sur les tokens inconnus du lexique de référence (en l'espèce, le L*efff*). Elle est complétée par un module de construction de candidats de correction pour les tokens connus mais identifiés comme susceptibles de pouvoir malgré tout résulter d'une altération. L'application d'un modèle de langue permet ensuite de prendre en compte le contexte des tokens pour lesquels des corrections candidates ont été proposées afin de choisir la ou les meilleure(s). Dans certains cas, la forme d'origine est conservée, et le modèle de langue peut donc choisir de ne pas la corriger. Ceci permet d'éviter de corriger un néologisme ou un emprunt qui n'aurait pas été reconnu comme tel (on se référera à la section 3.2 pour plus d'informations sur les techniques de détection et d'analyse des néologismes que nous avons développées). Toutefois, contrairement à ce qui se passe dans TEXT2DAG, aucune reconnaissance des mots composés n'est réalisée.

Notre technique de correction des tokens inconnus repose sur plusieurs hypothèses et approximations. Tout d'abord, nous avons constaté, dans les corpus que nous avions à traiter dans le cadre de ce travail, que la majorité des procédés d'altération correspondaient à l'application d'au plus une opération *lourde*, c'est-à-dire l'une des quatre opérations élémentaires sur laquelle repose la notion de distance d'édition, éventuellement associée à une altération légère (diacritique manquant ou incorrect). Nous limitons donc l'espace de recherche en conséquence. Par ailleurs, nous prenons en compte le fait que cette étape de correction est conçue pour faire suite à d'autres modules destinés à la détection des néologismes et des emprunts non adaptés (cf. section 3.2) ainsi qu'à la correction d'un certain nombre de types d'altérations formellement moins régulières, telles que *tjs* pour *toujours* (cf. notamment la section suivante). L'analogie est alors un mécanisme qui semble adapté à la tâche : ainsi, si l'on a déjà vu au préalable et extrait par

23. Le travail présenté dans cette section a été réalisé dans le cadre de la thèse de Marion Baranes dont j'étais encadrant principal et co-directeur et financée par l'entreprise viavoo, au sein de laquelle travaillait cette dernière. Avant la publication de la thèse proprement dite (Baranes, 2015), un état antérieur du travail a été décrit dans (Baranes et Sagot, 2014b).

une généralisation appropriée la règle de correction permettant de passer de « *engagemt* » à « *engagement* », il est possible de prédire l'orthographe correcte de « *changemt* ».

Nous avons appris nos règles de correction à partir du corpus de fautes lexicales WiCoPaCo, construit par Max et Wisniewski (2010) à partir des modifications apportées au fil des versions successives de la Wikipedia francophone, et dont nous avons extrait un ensemble de couples de tokens de la forme ⟨*token altéré*, *token bien orthographié*⟩ correspondant à des fautes lexicales [24]. Comme indiqué précédemment, nous faisons l'hypothèse que chaque token inconnu à traiter est le résultat d'au plus une altération « lourde ». Produire nos règles de correction consiste alors à extraire automatiquement de chaque couple de tokens du corpus d'apprentissage : (i) le contexte gauche complet précédant l'altération, (ii) l'altération et sa correction, (iii) le contexte droit complet suivant l'altération — les contextes gauche et droit complets sont identiques dans la forme altérée et dans la forme bien orthographiée [25], il est simple d'extraire les séquences de lettres constituant une faute et représentant la correction de cette faute. Ainsi, de la paire ⟨*souevnt*, *souvent*⟩ nous extrayons la *règle de base* notée `{#sou}{ev → ve}{nt#}`, où les « # » matérialisent le début et la fin du mot.

Une telle règle est bien trop détaillée, puisqu'elle ne s'applique qu'à `souevnt`. Elle doit donc être généralisée par comparaison avec les autres règles de base extraites. Afin d'éviter les cas de sous-correction et de sur-correction, nous construisons deux jeux de règles par deux types de généralisations différentes : un jeu de *règles spécifiques* et un jeu de *règles larges*, comme illustré au tableau 7.2 (cf. Baranes, 2015 pour des définitions précises). Nous avons également appliqué certaines généralisations à ces deux jeux de règles, au niveau de la règle de substitution elle-même, pour introduire un marqueur générique de doublement ou de dé-doublement de lettre, et pour représenter l'ajout, la suppression ou la substitution de diacritiques [26]. Enfin, toues nos règles sont pondérées à partir de la fréquence des règles de base qui leur ont donné naissance — à nouveau, on pourra se reporter à (Baranes, 2015) pour plus de détails.

24. Ce corpus est composé de 72 483 erreurs lexicales et de 74 100 erreurs grammaticales qui ont été annotées comme telles dans le corpus par leurs auteurs au moyen d'un processus automatique. Chacune des fautes est associée à sa correction, effectuée par un contributeur de la Wikipédia. Puisque dans ce travail nous nous intéressons uniquement aux fautes lexicales, nous n'avons utilisé que les fautes lexicales, c'est-à-dire les fautes annotées « *non_word_error* ». Par ailleurs, nous ne voulons pas que la fréquence d'un mot puisse biaiser la pondération des règles qui en seront extraites. Nous n'avons donc conservé qu'une seule occurrence de chaque faute annotée (soit un total de 36 344 fautes). Nous avons utilisé 60% de ces dernières, soit 21 581 fautes, comme données d'entraînement pour l'apprentissage de nos deux jeux de règles. Par ailleurs, environ 7,5% des fautes lexicales de WiCoPaCo (soit 2 731 fautes), sans recouvrement avec les données d'entraînement, ont été conservées pour constituer notre corpus de test.

25. Les règles sont apprises en ne prenant en compte que les paires impliquant une et une seule zone altérée. Les cas combinant par exemple une altération lourde et un changement de diacritique sont traitées *a posteriori* grâce aux règles apprises de la sorte.

26. D'où des règles comme `{Vt}{+_→_}{e#}` extraite de ⟨*fautte*, *faute*⟩, `{Vr}{_→+_}{eV}` de ⟨*ereur*, *erreur*⟩ ou `{Cr}{_→^}{eC}` de ⟨*arret*, *arrêt*⟩.

Règle	Poids	Exemple
Règles « spécifiques »		
{V[pfnlmctbsredg]}{_ → +_}{[aieuonyr]C}	0,970	**aten**dre → **atten**dre
{V[rmltdvcsnpyxg]}{a → e}{[nmilu]C}	0,660	r**alant**ir → r**alent**ir
Règles « larges »		
{C}{io → oi}{C}	0,298	**tiol**e → **toil**e
{[CV#]}{´ → ‘}{V}	0,738	é**lé**ve → é**lè**ve

Tableau 7.2 – Exemples de règles de correction extraites par analogie à partir des fautes lexicales contenues dans le corpus WiCoPaCo

Enfin, nous avons mis en place un système de correction élémentaire reposant sur la distance d'édition, en limitant le nombre d'opérations élémentaires possibles à une seule. Ceci permet de tenter de corriger des tokens inconnus sur lesquels aucune des règles précédemment apprises n'est applicable, mais a l'inconvénient d'être plus bruyant et de ne pas pouvoir être pondéré de la même façon. L'application de la correction générique est donc restreinte aux seuls tokens à corriger pour lesquels aucune règle extraite par analogie n'est applicable.

À ce stade, une première évaluation peut être réalisée, qui consiste à mesurer la qualité et le nombre de candidats ainsi proposés. En nous limitant aux seuls jeux de règles, seuls 5,3% des mots ne reçoivent aucun candidat. La correction générique permet de réduire ce taux à 3,6%[27]. La figure 7.2 détaille le nombre de candidats proposés pour chacun des tokens traités : on constate que ce nombre reste relativement bas dans la majorité des cas. Quant à la qualité de ces candidats, elle peut être estimée par l'évaluation d'un système complet qui choisirait systématiquement parmi les candidats proposés le candidat de correction correct s'il en fait partie ou, à défaut, un des candidats de normalisation corrects, ou encore, s'il n'y en a aucun, un candidat au hasard. Les résultats d'un tel système, obtenus sur nos données de test (cf. note 24), sont donnés au tableau 7.3 — on se reportera à (Baranes, 2015) pour plus d'informations. Ils constituent une borne supérieure pour le système réel qui fait usage d'un modèle de langue pour effectuer ces choix[28]. Ils justifient également l'emploi de la correction générique lorsque les règles apprises ne proposent aucun candidat.

Face à un token à traiter, toutes les règles (spécifiques et larges) qui peuvent être appliquées le sont. Une correction c proposée pour l'altération w est alors associée à trois poids : $S_s(c, w)$, poids de la règle spécifique permettant de passer de w à c, $S_l(c, w)$, poids

27. Les 93 mots non traités ont été étudiés manuellement : 85% d'entre eux correspondent à des fautes trop complexes ou trop nombreuses (ex : *arondisemernts*, *aérdorme*) et 15% à des séquences de caractères difficilement interprétables (ex : *klàkoes*, *piwut*).

28. Si l'on remplace l'oracle par un choix aléatoire parmi les candidats proposés, on obtient une borne inférieure qui, pour la f-mesure, est d'environ 57% en correction et 68% en normalisation (la correction générique ne faisant là aussi gagner qu'environ un demi pourcent.

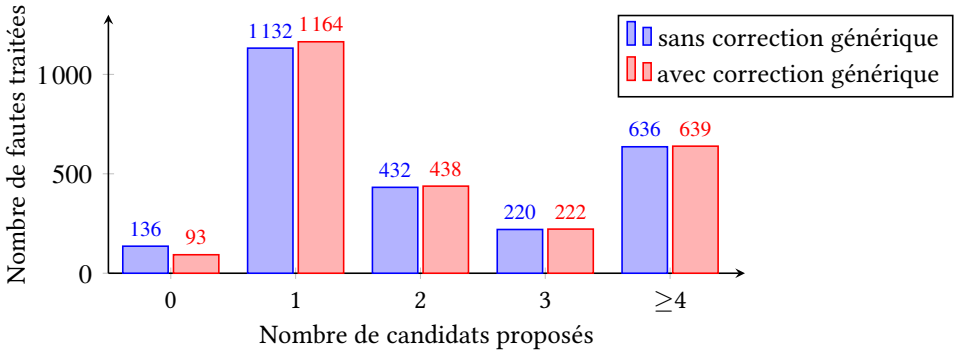


FIGURE 7.2 – Nombre de candidats par token inconnu traité, avec et sans activation du module de correction générique.

	SANS CORRECTION GÉNÉRIQUE		AVEC CORRECTION GÉNÉRIQUE	
	CORRECTION	NORMALISATION	CORRECTION	NORMALISATION
PRÉCISION	94,1	95,6	93,7	95,1
RAPPEL	89,1	90,4	90,3	91,6
F-MESURE	91,6	92,9	92,0	93,3

TABLEAU 7.3 – Évaluation des candidats de correction produits par l'utilisation des règles extraites par analogie. Les résultats indiqués sont obtenus par l'évaluation de la qualité de la correction (resp. normalisation) que l'on obtiendrait si un oracle choisissait, pour chaque token altéré, la normalisation (resp. correction) correcte si elle fait partie des candidats.

de la règle spécifique permettant de passer de w à c, et $F(c)$, fréquence de c en corpus normalisée entre 0 et 1. Le coût de la correction de w en c est alors une moyenne pondérée de ces trois poids, dont les coefficients ont été choisis après de nombreuses expériences décrites dans (Baranes et Sagot, 2014b) et plus en détails dans (Baranes, 2015).

Nous ne détaillerons pas ici la façon dont ce système est complété par un module d'ajout de candidats de normalisation à certains tokens connus du lexique de référence, pour tenter de corriger les altérations résultant de fautes grammaticales. Nous ne détaillerons pas non plus la mise en place d'un modèle de langue pour désambiguïser totalement ou partiellement le DAG complet de possibilités que l'on obtient alors. Le lecteur intéressé pourra se reporter à (Baranes, 2015). Les résultats du système complet de normalisation sur des données réelles traitées par l'entreprise viavoo sont toutefois donnés dans le tableau 7.4. De nombreuses pistes restent naturellement à creuser pour améliorer ce système, dont des pistes étudiées dans d'autres travaux (passage intermédiaire par la

	Corpus représentatif		Corpus très bruité	
	1-best	*3-best*	*1-best*	*3-best*
Préc.	76,1	80,5	74,1	80,2
Rapp.	54,7	59,8	58,3	62,3
F-mes.	63,6	68,6	65,3	70,1

Tableau 7.4 – Évaluation du système complet de normalisation de textes bruités sur deux jeux de données réelles : un corpus représentatif des données traitées par l'entreprise viavoo (13 896 tokens dont 250 altérations non flexionnelles) et un corpus de données fortement bruitées rassemblé manuellement par l'entreprise (1749 tokens dont près de 200 altérations non flexionnelles).

forme phonétisée, notamment). Plus problématique est l'utilisation d'un modèle de langue appris sur des textes réels alors que l'on cherche ici à normaliser et non à corriger [29].

7.3.3 Correction lexicale déterministe par règles opérant au niveau des tokens et développées manuellement

Une autre solution, plus simple quoique moins générique, peut permettre de traiter des altérations bien plus fortes et surtout moins régulières que celles que l'on peut traiter par les méthodes décrites aux deux précédentes sections. Il s'agit d'altérations que l'on peut notamment trouver dans les productions des utilisateurs du web (*user-generated content*), dont voici quelques exemples attestés issus du French Social Media Bank, corpus arboré de données issues du web et sur lequel nous reviendrons notamment à la section 8.4.1.

> Je soupçonnes que "l'enfarineuse" était en faite une cocaineuse vu la pêche de #Hollande ce soir à #Rouen. [Twitter]
>
> Ces pas possible déjà que battelfield a un passe online [JeuxVideo.com]
>
> Si y'a que Juliet &Zayn qui sont co' sur le RPG,et qui font leur vie tranquilles [JeuxVideo.com]
>
> car je ne me senté pa desiré, pa aimé, pa bel du cou, g t pa grd chose en fet. [Doctissimo.fr]
>
> L'Ange Michael vraiment super conten pour toi mé tora plus grace a moi tkt love you ! [FaceBook]
>
> @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh [Twitter]

Cette solution consiste simplement à disposer d'une liste de règles de réécriture à appliquer de façon déterministe, typiquement par exemple par ordre décroissant de

29. Répondre correctement à cette difficulté ne va cependant pas de soi. La seule solution pour disposer d'un corpus permettant d'apprendre un modèle de langue adapté à la normalisation de textes serait de partir d'un corpus édité et de changer aléatoirement certaines formes en d'autres tout en restant à chaque fois parmi les formes du même lemme. Ainsi, on disposerait d'un corpus ressemblant à ce que pourrait produire un système de normalisation parfait. Mais la proportion et la distribution de ces altérations aléatoires reste à définir.

longueur pour traiter les cas particuliers avant les cas plus généraux. On dispose alors d'un correcteur non ambigu, dont la sortie pourra être des « tokens corrigés » ou directement des formes. Cela suppose néanmoins le développement d'une liste de règles, et la méthode est donc à la fois moins générique et plus coûteuse en temps de travail manuel que les méthodes précédentes. Néanmoins, il est possible de traiter rapidement une proportion importante des altérations les plus fortes en raison de leur fréquence relativement élevée. Par ailleurs, la correspondance entre tokens d'entrée et tokens ou formes de sortie doit être préservée, quand bien même elle n'est pas bijective.

C'est cette stratégie que nous avons développée pour le traitement de corpus issus du web en anglais et en français, en nous contentant de règles produisant des « tokens corrigés ». Nous avons complété cette stratégie, pour les tokens inconnus plus longs qu'un certain seuil, par une recherche dans le lexique aux diacritiques près, recherche dont le résultat n'est exploité que si elle renvoie exactement une réponse [30].

À ce jour, la principale utilisation que nous avons faite de cette méthodologie de correction consiste à l'utiliser pour débruiter temporairement des textes à étiqueter morphosyntaxiquement avec notre étiqueteur statistique MElt, sur lequel nous reviendrons aux sections 8.1 et 8.4. L'idée est qu'un étiqueteur comme MElt, entraîné sur du texte edité, aura de meilleures performances sur du texte qui ressemble à du texte édité. Nous reviendrons à la section 8.4 sur cette méthodologie d'étiquetage morphosyntaxique et sur les résultats quantitatifs que nous avons obtenus tant pour le français [31] que pour l'anglais [32].

En anglais, notre objectif était de pouvoir pré-traiter les corpus de la campagne SANCL 2012 d'évaluation des analyseurs syntaxiques pour l'anglais sur des textes issus du web (Petrov et McDonald, 2012), afin d'améliorer leur étiquetage morphosyntaxique. Nous reviendrons à la section 8.4 sur cette méthodologie d'étiquetage morphosyntaxique couplé à la correction par règles, à la section 8.4.3 sur les résultats que nous y avons obtenus en étiquetage morphosyntaxique et à la section 9.4.2 sur nos résultats en analyse en constitants (Seddah *et al.*, 2012b), lesquels nous ont permis d'arriver classés deuxièmes lors de cette campagne.

Pour le français, nous avons utilisé cette même technique de couplage entre correction et étiquetage morphosyntaxique d'une façon différente, puisqu'elle nous a servi à pré-annoter certaines parties du French Social Media Bank en préparation au travail de

30. Voir l'exemple du token `preliminaires` dans la table 7.6.

31. Un premier jeu de règles pour le français avait été construit à partir de séquences de tokens candidates extraites automatiquement des données brutes du French Social Media Bank, séquences auxquelles nous avons associé manuellement une correction si elle était nécessaire et pouvait être faite de façon déterministe. Ceci nous a permis le développement de plusieurs centaines de règles. Ce chiffre atteint aujourd'hui 1872 règles, notamment grâce à des travaux préliminaires sur certaines parties du corpus CoMéRé, corpus de données françaises médiées par les réseaux et développé sous la houlette de Thierry Chanier, dans le cadre du Consortium Corpus Écrits de la TGIR Huma-Num (Chanier *et al.*, 2014).

32. La même approche que celle appliquée pour le français sur le French Social Media Bank nous a permis, à partir des données du Google Web Treebank, d'extraire 327 règles de réécriture pour l'anglais.

validation et de correction par des annotateurs humains. Le French Social Media Bank (FSMB) est un corpus arboré de données produites par les utilisateurs de l'internet (Seddah *et al.*, 2012c,d). Ce corpus, le premier de ce type pour une langue autre que l'anglais, a été constitué après une étude détaillée des phénomènes linguistiques et typographiques à l'œuvre dans des données de ce type (cf. tableau 7.1 ainsi que Seddah *et al.*, 2012c et Baranes, 2015). Une métrique dédiée, que nous avons développée à partir de la divergence de Kullback-Leibler entre les distributions des caractères dans deux corpus distincts, nous a permis d'identifier deux sous-ensembles de données : les donnés faiblement bruitées et les données fortement bruitées. C'est sur ces dernières que nous avons couplé correction et étiquetage, les autres sous-corpus ayant été directement pré-annotés par la version standard de MElt ($\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$). Nous avons néanmoins conservé certains sous-corpus fortement bruités comme corpus de test, et nous présenterons à la section 8.4.2 nos différents résultats en étiquetage morphosyntaxique sur le FSMB — nos résultats en analyse syntaxique en constituants sont donnés à la section 9.4.1.

On note donc une différence cruciale entre ces travaux pour le français et les travaux pour l'anglais dans le contexte de la campagne SANCL 2012 : pour le développement du FSMB, les corpus à corriger nous étaient connus à l'avance, et le développement de règles de correction à partir du corpus à traiter était légitime, l'objectif étant d'accélérer l'annotation manuelle qui suivait ; pour la campagne SANCL, les corpus à corriger nous étaient naturellement inconnus, et nos règles ont été développées à partir d'autres données. Nous verrons à la section 8.4.1 l'impact de ces corrections sur la qualité de l'étiquetage morphosyntaxique automatique que nous avons réalisé dans les deux cas, et la façon dont nous avons couplé correction et étiquetage.

Les tableaux 7.5 et 7.6 illustrent le résultat de l'application de jeux de règles de correction, mais aussi de certaines grammaires locales de SxPipe, sur deux exemples : un exemple forgé en anglais et un exemple réel en français tiré du FSMB [33].

33. Comme l'illustre le tableau 7.5, les règles de correction peuvent prendre la forme d'expressions régulières. Nous avons par exemple en français la règle « `dois ([^ ]{2,})é` → `dois $1er` », qui indique qu'un token en `-é` suivant le token `dois` doit voir son `-é` final remplacé par `-er`. Par ailleurs, le nombre de « tokens corrigés » peut être différent du nombre de tokens de départ ; Dans ce cas, nous utilisons des correspondances 1 à *n* ou *n* à 1. Par exemple, la règle `ni a pa` → `n' y a pas` indique explicitement au moyen du symbole « ␣ » que le token `ni` est ici à remplacer par les deux tokens `n'` et `y`, alors que `pas` est la correction du seul token initial `pa` [34]. Un autre exemple serait celui de la séquence `c t` (pour *c'était*), dans laquelle `c` correspond à la fois à *c'* et à la première syllabe de *était*. Autrement dit, il faut considérer `c t` comme une transcription sous la forme d'un composé de ce qui est fondamentalement l'amalgame de *c'* et *était*. La règle s'écrit donc « `c t` → `c'était` ».

Tokens	« Tokens corrigés »	Règle appliquée
i	I	i → I
know	know	
im	I 'm	im → I␣'m
gon na	going to	gon na → going to
visit	visit	
somewebsite.com	*_URL*	*Grammaire SxPipe de reconnaissance des URL*
! ! ! ! !	!	([!?])+ → \1

Tableau 7.5 – Tokens de départ et tokens normalisés par l'application du jeu de règles développé manuellement et intégré à MElt pour l'anglais, sur la phrase d'exemple *i know im gon na visit somewebsite.com* 'je sais ke jvé aller sur somewebsite.com'. L'exemple est forgé mais permet d'illustrer différents phénomènes, y compris le résultat de l'application préalable de grammaires locales, qui ne produisent pas des « tokens corrigés » mais bien des formes spéciales, comme décrit plus haut. La tokenisation de départ n'est pas fournie par SxPipe mais reproduit la tokenisation fournie par les organisateurs de la tâche SANCL 2012 (cf. sections 8.4.3 et 9.4.2).

Tokens	« Tokens corrigés »	Règle appliquée
sa fé	ça fait	sa fé → ça fait
o moin	au moins	o moins → au moins
6 mois	6 mois	
qe	que	qe → que
les	les	
preliminaires	préliminaires	*Recherche à l'accent près pour les mots longs*
sont	sont	
" sauté "	" sauté "	
c a dire	c'est-à-dire	c␣a␣dire → c'est-à-dire
qil	qu' il	qil → qu'␣il
yen	y en	yen → y␣en
a	a	
presk	presque	presk → presque
pa	pas	pa → pas

Tableau 7.6 – Tokens de départ et tokens normalisés par l'application du jeu de règles développé manuellement et intégré à MElt pour le français, sur la phrase d'exemple *sa fé o moin 6 mois qe les preliminaires sont "sauté" c a dire qil yen a presk pa.* L'exemple est extrait du sous-corpus fortement bruité de la section *Doctissimo* du French Social Media Bank.

7.3.4 Correction déterministe au sein d'entités typographiques [35]

Les entités typographiques n'échappent pas au problème des erreurs orthographiques ou typographiques, notamment dans le cas des systèmes de reconnaissance optique des caractères (OCR). Cependant, les erreurs au sein des entités typographiques sont difficiles à corriger dans la mesure où il est pour ainsi dire impossible de disposer d'un lexique des entités valides ou d'un lexique de corrections déterministes comme à la section précédente, sauf pour les cas les plus fréquents. À l'inverse, il y a un avantage à ce que ces entités soient régies par des règles de bonne formation spécifiques au type d'entité. Si par exemple, au sein de la sortie d'un système d'OCR, on reconnaît dans le token `A1203` une formule chimique, notamment en raison de son contexte, il pourra être corrigé en `Al2O3` voire en Al_2O_3. Mais ce même token, s'il n'est pas identifié comme étant une formule chimique, sera conservé tel quel, avec une bonne chance qu'il s'agisse de la bonne solution.

Nous avons donc amélioré la robustesse des grammaires locales de la chaîne SxPipe (adresses…), et notamment celle des nouvelles grammaires mentionnées précédemment (formules chimiques…) et développé un module de correction travaillant de façon différenciée sur certains types d'entités nommées fréquemment rencontrées dans des corpus spécialisés. Contrairement à une approche statistique reposant par exemple sur des modèles d'erreurs, cette approche par règles permet une meilleure prise en charge des propriétés formelles non linguistiques des entités. De plus, les règles de corrections ne sont pas seulement spécifiques à chaque type d'entité, elles le sont parfois à des types de segments au sein des entités (par exemple, le numéro d'arrondissement dans une adresse). Une évaluation manuelle des résultats de ces corrections pour trois types d'entités différentes a été réalisée. Elle appuie sur des corpus d'écart de qualité éditoriale produits par l'entreprise Numen Digital, données numérisés automatiquement puis corrigées manuellement, ce qui fournit après réalignement automatique un jeu de données d'évaluation [36]. Les résultats sont donnés à la table 7.7. Pour plus de détails sur les corpus et la méthodologie d'évaluation, on pourra se reporter à (Sagot et Gábor, 2014). Ces résultats montrent la pertinence d'une telle approche, au moins dans le cas de la correction de sorties d'OCR.

35. Le travail évoqué dans cette section a été réalisé en collaboration avec Kata Gábor dans le cadre du projet PACTE, consacré à la correction automatique de sorties de systèmes de reconnaissance optique de caractères, ou OCR. Il a donné lieu à deux publications (Gábor et Sagot, 2014 ; Sagot et Gábor, 2014) et plusieurs rapports techniques.

36. Les dates et adresses sont évaluées sur des données de la collection Gallica fournies par la Bibliothèque Nationale de France. Les formules chimiques le sont sur un corpus de demandes de brevets déposées à l'Agence Européenne des Brevets.

37. Ce chiffre élevé s'explique par le fait que de très nombreuses formules chimiques fassent l'objet d'une mise en indice (parfois en exposant) des nombres qu'elles contiennent (ainsi, `H20` est corrigé en H_2O. Toutefois, de nombreuses autres modifications sont également effectuées, qui concernent des erreurs plus critiques.

Type d'entité	Reconnaissance			Correction/normalisation		
	Préc.	Rapp.	#occ. pour 10^6 tokens	Préc.	Rapp.	#entités corrigées /#entités détectées
Dates	0,98	0,97	4713	0,96	–	2%
Adresses	0,83	0,86	269	0,76	–	3%
Formules chimiques	0,91	(0,88)	300	0,95	0,90	72% [37]

Tableau 7.7 – Evaluation sur des sorties d'OCR de la reconnaissance robuste de certains types d'entités nommées par certaines grammaires locales de SxPipe et de leur correction/normalisation par le module dédié développé au cours du projet PACTE.

7.4 Composés : le cas de l'identification des formes en mandarin [38]

L'identification des formes composées à partir des tokens — tâche à ne pas confondre avec l'identification des mots sémantiques, bien plus délicate — peut donc être effectuée sur des textes utilisant un système d'écriture à séparateur typographique, y compris lorsque certains tokens sont altérés. Bien que non triviale, cette tâche bénéficie du fait que, dans ce type de système d'écriture, les tokens et les formes, bien que n'entretenant pas systématiquement une relation bijective, se correspondent souvent.

La granularité du découpage en tokens est différente lorsque l'on traite de corpus produits dans un système d'écriture faisant usage de sinogrammes. Comme nous l'avons évoqué à la section 1.3.7, la définition la plus simple consiste alors à considérer comme un token chaque sinogramme ainsi que de chaque caractère utilisé en complément par des systèmes d'écriture comme celui du japonais (hiragana, katakana). Regrouper en formes d'une part des sinogrammes et d'autre part des tokens du français n'est donc pas exactement la même tâche. Dans cette section, nous allons nous pencher brièvement sur le mandarin, pour lequel il est d'usage de nommer *segmentation* la tâche consistant à regrouper les sinogrammes en formes [39].

La tâche dont il sera question dans cette section est donc inverse de celle qu'il faut mettre en œuvre pour traiter des systèmes d'écritures où les tokens regroupent typiquement plusieurs formes, comme c'est le cas notamment en sanskrit. Cette langue est en effet l'exemple le plus usuel des langues au sein desquelles opèrent des mécanismes morphophonologiques ou morphographémiques (généralement appelés *sandhi* dans le cas

38. Cette section donne un aperçu des travaux que j'ai menés en collaboration avec Pierre Magistry au cours de sa thèse dont j'étais l'un des co-directeurs. Il a donné lieu, outre la thèse elle-même (Magistry, 2013), à plusieurs publications (Magistry et Sagot, 2011 ; Magistry, 2012 ; Magistry et Sagot, 2012, 2013).

39. Le système que nous présentons ici n'est pas spécifique au mandarin, ni même à l'identification de formes à partir de séquence de sinogrammes ou d'autres types de tokens comparables (« mots » du vietnamien, par exemple). Des expérimentations préliminaires sur d'autres langues et à partir d'autres unités initiales (lettres, phonèmes, tokens) ont été réalisées, toujours pour tenter d'obtenir des regroupements en unités plus large, typiquement en formes. Toutefois, nous ne sommes en mesure de fournir d'évaluation complète que pour le mandarin, et nous nous limitons donc ici à cette langue.

du sanskrit), lesquelles altèrent les formes, notamment à la frontière entre deux formes (*sandhi externe*). De plus, le système d'écriture utilisé aujourd'hui habituellement pour transcrire le sanskrit est une variante du système d'écriture *devanāgarī* dans laquelle les composés et les suites de formes entre lesquelles ont lieu des phénomènes de sandhi ne sont pas séparés par des espaces. Cela conduit à de nombreux tokens multi-formes, dans lesquelles il est donc utile de savoir identifier la séquence de formes sous-jacente, séquence appelée *padapāṭha* par les sanskritistes [40]. Des concepts, des outils et des ressources informatiques adaptés à cette tâche ont notamment été développés par Huet (2005, 2009).

Un grand nombre de méthodes ont été proposées pour effectuer une segmentation automatique du mandarin. Certaines reposent sur des règles et des lexiques, d'autres utilisent des méthodes d'apprentissage automatique supervisé ou non supervisé. Six campagnes du « *Chinese Word Segmentation Bakeoff* » ont été organisées par l'ACL. Zhao et Liu (2010) donnent un résumé des performances obtenues par les systèmes en compétition lors de la campagne de 2010. Ils soulignent que, si la précision peut sembler satisfaisante, la tolérance au changement de domaine et la reconnaissance des mots inconnus restaient les limitations majeures. Notons que lors de cette campagne, le système de base (*baseline*) et le meilleur système (*topline*) sont obtenus avec le même algorithme, un simple *maximum-matching* (*minimisation du nombre de mots*) reposant sur un inventaire d'unités lexicales, et ne se distinguent que par le lexique utilisé : la *baseline* utilise un lexique extrait à partir du corpus d'entraînement, tandis que la *topline* utilise un lexique extrait à partir de la totalité du corpus et connaît donc toutes les formes attendues. Xue (2003) commente les résultats d'une autre heuristique simple qui repose sur un lexique, celle dite du *longest-match* gauche-droite (*plus longue chaîne d'abord*) : cette heuristique fournit de très bons résultats (f-mesure de 0,952) si le lexique est exhaustif mais se dégrade très rapidement lorsque le corpus de test contient des mots inconnus (f-mesure de 0,898). Le *maximum-matching* utilisé lors du *bakeoff* obtient quant à lui des scores (f-mesure) supérieurs à 0,98 sur différents corpus (la *topline*) avec un lexique exhaustif et des scores de 0,72 à 0,88 selon les domaines dans la configuration *baseline*. Les 18 systèmes présentés lors du *bakeoff* de 2010 ont tous obtenu des résultats intermédiaires entre ces deux niveaux. Il faut donc souligner l'importance pour cette tâche des ressources lexicales.

De nombreux systèmes de segmentation par apprentissage supervisé du mandarin ont été proposés mais ils requièrent des corpus segmentés manuellement. Ceux-ci sont souvent spécifiques à un genre, un domaine ou une variété de mandarin et en l'absence d'un consensus sur la définition de ce qu'est un « mot », ils suivent des guides

40. Huet (2009) reprend un exemple du grammairien Patañjali : soit le token `śvetodhāvati` (en translittération latine usuelle). Il peut être analysé comme recouvrant la séquence de deux formes *śvetaḥ dhāvati* LE [CHEVAL…] BLANC COURT (LITTÉRALEMENT BLANC+COURT) *via* la règle de sandhi *-aḥ dh-* → *-odh-*. Mais il peut être également compris comme correspondant à la séquence de trois formes *śvā itaḥ dhāvati* LE CHIEN COURT VERS ICI (LITTÉRALEMENT CHIEN+(VERS) ICI+COURT), où la règle de sandhi *-ā i-* → *-e-* est appliquée en plus de la précédente.

d'annotations qui divergent (Kratochvíl, 1967 ; Huang, 1984 ; Duanmu, 1998 ; Packard, 2000 ; Magistry, 2013). Les systèmes supervisés atteignent aujourd'hui des résultats satisfaisants lorsque le corpus approprié pour l'entraînement est disponible. Cependant, si l'on veut faire face à une plus grande diversité en genres et en domaines ou répondre à des questions plus théorique sur la caractérisation formelle des unités de langue, s'intéresser aux approches non supervisées nous semble nécessaire. C'est donc la piste que nous avons investiguée. De telles approches donnent lieu, en effet, à la définition de mesures quantitatives objectives permettant de faire émerger une notion de « mot » qui, pour n'être linguistiquement que partiellement pertinente [41], n'en a pas moins l'avantage d'être définie d'une façon implémentable et reproductible.

Notons toutefois que ces méthodes sont plus difficiles à évaluer que les méthodes supervisées, puisqu'il n'existe pas *a priori* de raison pour que la sortie d'un tel système corresponde à un guide de segmentation plutôt qu'à un autre. Toutes nos évaluations sont donc effectuées successivement sur chacun des quatre principaux corpus segmentés manuellement disponibles pour le mandarin, ceux de la campagne d'évaluation de 2005 (*2005 Chinese Word Segmentation Bakeoff*, Emerson, 2005). Il s'agit du corpus de l'Université de Pékin, qui suit le guide de Yu *et al.* (2002a,b), du corpus équilibré de l'Academia Sinica, qui suit Huang *et al.* (1996), du corpus de la Hong-Kong City-University, extrait du corpus LIVAC (T'sou *et al.*, 1997) et du corpus de Microsoft Research. L'interprétation des résultats que nous donnerons par la suite nécessite de disposer de l'information suivante : si l'on entraîne des segmenteurs supervisés sur chacun des corpus, ils obtiennent des scores de l'ordre de 95% à 97% sur des données annotées selon les mêmes conventions ; mais évalués sur les trois autres corpus, leurs scores tombent tous en dessous des 90%, et descendent jusqu'à 81% dans certaines configurations (Magistry, 2013, ch. 7). C'est dire si les différences entre les conventions sous-jacentes sont importantes, et explicite le caractère peu universel de la notion de « mot » en linguistique chinoise.

7.4.1 Segmentation non supervisée reposant sur la variation de l'entropie de branchement

Une des approches les plus utilisées pour la segmentation non supervisée du mandarin est insipirée par une hypothèse linguistique formulée par Harris (1955). Reformulée au moyen de la notion d'entropie de branchement (*Branching Entropy*, *BE*) par Tanaka-Ishii (2005) en suivant les travaux de Kempe (1999), cette hypothèse peut s'énoncer comme suit : si les séquences de graphèmes, phonèmes, ou autres produites par l'homme étaient aléatoires, on s'attendrait à ce que l'entropie de branchement d'une séquence (estimée à partir de *n*-grammes en corpus) décroisse lorsque la longueur de la séquence croît.

41. Notamment en l'absence de prise en compte de la notion de classe de mots, notion que l'on peut penser à capturer par des approches distributionelles ou analogiques, et qu'il conviendrait de coupler avec le modèle de segmentation.

Ainsi, la variation de l'*entropie de branchement* (*Variation of the Branching Entropy, VBE*) devrait être systématiquement négative. Lorsque l'on observe au contraire une VBE positive, l'hypothèse de Harris conduit à conclure que l'on se situe à une frontière entre unités linguistiques. Plus formellement, on peut définir l'entropie de branchement comme suit : soit n-gramme $w_{0..n} = w_{0..1}\, w_{1..2} \ldots w_{n-1..n}$ de contexte droite $\chi_{\rightarrow}$, son *entropie de branchement à droite* $h_{\rightarrow}(w_{0..n})$ est définie par

$$\begin{aligned} h_{\rightarrow}(w_{0..n}) &= H(\chi_{\rightarrow} \mid w_{0..n}) \\ &= -\sum_{w \in \chi_{\rightarrow}} P(w \mid w_{0..n}) \log P(w \mid w_{0..n}). \end{aligned}$$

L'*entropie de branchement à gauche* est définie de façon symétrique par $h_{\leftarrow}(w_{0..n}) = H(\chi_{\leftarrow} \mid w_{0..n})$. La variation de l'entropie de branchement de $w_{0..n}$, ou VBE, est alors définie pour chaque direction comme suit :

$$\begin{aligned} \delta h_{\rightarrow}(w_{0..n}) &= h_{\rightarrow}(w_{0..n}) - h_{\rightarrow}(w_{0..n-1}) \\ \delta h_{\leftarrow}(w_{0..n}) &= h_{\leftarrow}(w_{0..n}) - h_{\leftarrow}(w_{1..n}). \end{aligned}$$

Nous avons montré qu'il était alors pertinent de normaliser ces VBE en fonction de la longueur des formes en nombre de sinogrammes. On se reportera par exemple à (Magistry, 2013, ch. 8) pour plus de détails sur cette étape de normalisation. Nous utiliserons désormais ces VBE normalisées, notées respectivement $\tilde{\delta} h_{\rightarrow}(w_{0..n})$ et $\tilde{\delta} h_{\leftarrow}(w_{0..n})$ [42]

Nous avons tout d'abord proposé un système qui repose uniquement sur la VBE normalisée, en définissant une mesure d'*autonomie* définie sur les séquences de tokens (Magistry et Sagot, 2012). L'autonomie d'une séquence de tokens, c'est-à-dire d'un candidat-forme x, y est définie par $a(x) = \tilde{\delta} h_{\leftarrow}(x) + \tilde{\delta} h_{\rightarrow}(x)$. Cette fonction d'autonomie est alors utilisée par un algorithme de segmentation qui maximise la somme des autonomies des mots de la phrase à traiter. La segmentation $\hat{W}(s)$ choisie parmi l'ensemble $Seg(s)$ de toutes les segmentations possibles pour une phrase s donnée est donc définie

42. Plusieurs algorithmes de segmentation non supervisés qui reposent sur l'entropie de branchement ou sur la VBE ont été proposés. Cohen *et al.* (2002) utilisent l'entropie de branchement comme indicateur dans leur système par vote d'experts. Ils pointent la nécessité d'une normalisation mais utilisent directement l'entropie de branchement et non la VBE. Jin et Tanaka-Ishii (2006) proposent un systeme utilisant la VBE et l'évaluent par rapport à un corpus de mandarin segmenté manuellement. Zhikov *et al.* (2010) utilisent l'entropie de branchement pour obtenir une segmentation initale, en positionnant une frontière entre formes dès lors que l'entropie de branchement dépasse un certain seuil, lequel est déterminé en faisant appel à la notion de Longueur de Description (cf. ci-dessous). Wang *et al.* (2011) proposent l'algorithme ESA (*Evaluation, Selection and Adjustment*), un système plus complexe qui combine itérativement une mesure de cohésion et une mesure de non-cohésion, toutes deux reposant sur l'entropie de branchement. Ils obtiennent des résultats qui étaient à l'état de l'art au moment de la finalisation de ces expériences (fin 2013) mais, comme nous allons le voir, leur méthode n'est pas strictement non supervisée.

comme suit :

$$\hat{W} = \underset{W \in Seg(s)}{\arg\max} \sum_{w_i \in W} a(w_i) \cdot len(w_i).$$

Nous avons alors comparé les résultats de notre système à ceux de Wang *et al.* (2011). Ils sont légèrement inférieurs aux leurs, mais leur système a deux inconvénients par rapport au nôtre, dont le second est rédhibitoire : (i) il est beaucoup plus complexe, et surtout (ii) contrairement à ce qu'affirement leurs auteurs, il n'est pas non supervisé, puisqu'un paramètre crucial de leur modèle, dont la valeur a un impact important sur les résultats, est fixé par optimisation des scores sur un corpus segmenté manuellement.

7.4.2 Raffinement par minimisation de la longueur de description

Une autre approche fréquemment utilisée en segmentation du mandarin repose sur la notion de Longueur de Description, notion que nous avons déjà utilisée à plusieurs reprises, notamment au chapitre 4. L'idée sous-jacente est qu'un corpus segmenté peut être encodé sous la forme d'un lexique et d'une suite d'identifiants d'entrées lexicales dans ce lexique. Une bonne segmentation devrait conduire à une meilleure compacité de cet encodage par rapport à des segmentations moins pertinentes. Plus formellement, notons L le lexique, qui code chaque mot avec un code dont la longueur dépend de l'entropie du mot. Un corpus m, qui joue le rôle d'un message à transmettre, sera codé comme la séquence $C_L(m)$ de codes de L et aura pour longueur de description $DL_L(m) = DL(C_L(m), L) = DL(C_L(m)) + DL(L)$. Le lexique est lui-même supposé codé de façon optimale à partir des entropies des caractères qui le composent, caractères pris dans un alphabet noté S. Les longueurs des codes sont calculées comme au chapitre 4 à partir de l'entropie de l'unité qu'elles codent. Un objectif possible est alors de trouver la segmentation qui induit un lexique minimisant la valeur de $DL_L(m)$: c'est le principe de la longueur de description minimale, ou MDL (pour *Miminum Description Length*). En segmentation non supervisée du mandarin, la MDL est souvent utilisée de l'une des deux façons suivantes : elle aide à optimiser de façon non supervisée la valeur d'un paramètre du modèle (Zhikov *et al.*, 2010 ; Hewlett et Cohen, 2011) ou elle est directement utilisée comme mentionné ci-dessus, c'est-à-dire pour trouver une segmentation induisant une longueur de description minimale. Reste qu'il est impossible en pratique de chercher directement la segmentation minimisant la longueur de description, et il faut en ressortir à des heuristiques qui permettent de limiter l'espace de recherche. Zhikov *et al.* (2010) proposent deux techinques distinctes pour ce faire, qui ont toutes deux des inconvénients (Magistry et Sagot, 2013).

Nous avons donc proposé une nouvelle approche, qui utilise l'algorithme présenté ci-dessus (Magistry et Sagot, 2012) comme initialisation, puis cherche à minimiser la longueur de description au moyen d'heuristiques évaluant la pertinence de changements consistant à rajouter ou à supprimer des frontières entre formes à chaque position

inter-sinogramme (Magistry et Sagot, 2013). Ces heuristiques reposent sur la notion d'autonomie définie ci-dessus. Elles ont sur les heuristiques de Zhikov *et al.* (2010) l'avantage de permettre l'évaluation simultanée de l'effet d'un grand nombre de changements tout en restant au niveau du corpus et non du lexique, ce qui permet par exemple la création de formes longues précédemment absentes du lexique.

À ce stade, nous avons obtenu le résultat contre-intuitif suivant : les longueurs de description obtenues étaient inférieures à celles de Zhikov *et al.* (2010), qui disposent pourtant d'une étape d'optimisation de la longueur de description, et pourtant nos résultats en segmentation étaient moins bons que les leurs et même significativement moins bons que ceux obtenus par notre seul algorithme de base, celui de Magistry et Sagot (2012), utilisé ici comme initialisation (cf. tableau 7.8). Ainsi, avec les heuristiques dont nous disposons pour approcher la MDL, minimiser la longueur de description n'est finalement pas le meilleur moyen d'identifier les unités élémentaires [43]. Restait à savoir s'il y a un moyen de tirer malgré tout parti de l'intuition sous-jacente à la MDL. Nous avons donc réalisé une rapide analyse d'erreur. Il s'est avéré que les erreurs faites par l'étape de MDL étaient principalement de trois ordres : (i) fusionner des entités nommées considérées dans les corpus de référence comme composées de plusieurs formes, (ii) fusionner des mots fonctionnels mono-sinogramme avec des mots pleins avec lesquels ils cooccurrent fréquemment, et (iii) séparer en deux des formes bigrammes correctes produites par l'étape d'initialisation.

Nous avons donc proposé de contraindre l'étape utilisant la MDL au moyen de critères simples mais, il est vrai, dépendants de la langue. Ces critères sont les suivants, qui correspondent bijectivement à chacun des types d'erreurs identifiés ci-dessus : (i) aucune forme de plus de trois sinogrammes ne pourra être créée [44], (ii) nous interdisons la fusion des mots fonctionnels mono-sinogramme appartenant à une liste close [45], et (iii) nous interdisons la séparation des bigrammes en deux unigrammes, les premiers étant bien plus fréquents dans le lexique que les derniers (quand bien même les mots mono-sinogrammes sont fréquents en corpus). Comme le montre le tableau 7.8, nos résultats ont alors dépassé l'état de l'art en segmentation non supervisée du mandarin [46].

7.4.3 Amélioration par détection des entités typographiques

Le lecteur intéressé par des analyses qualitatives des erreurs produites par nos systèmes pourra se reporter à (Magistry, 2013). Un des points intéressants est toutefois le suivant :

43. La segmentation de référence reste systématiquement celle qui conduit à la longueur de description la plus courte.

44. Il existe en mandarin des formes composées de quatre sinogrammes, mais elles sont très rares.

45. Nous avons utilisé la liste suivante : 的 , 了 , 上 , 在 , 下 , 中 , 是 , 有 , 和 , 与 , 和 , 就 , 多 , 于 , 很 , 才 , 跟 .

46. Les scores donnés pour la méthode de Zhikov *et al.* (2010) sont le fait d'une réimplémentation de cette méthode par Pierre Magistry, afin de pouvoir obtenir des résultats sur les quatre corpus d'évaluation.

Méthode	**F-score**	**Longueur de description** (Mb)
Corpus de l'Université de Pékin		
Zhikov *et al.* (sans MDL)	0,719	17,9
Zhikov *et al.* (avec leur MDL)	0,808	15,6
Notre système (sans notre étape MDL)	0,786	16,1
Notre système (avec notre étape MDL non contrainte)	0,729	**15,2**
Notre système (avec notre étape MDL contrainte)	**0,832**	15,6
Gold	*1,0*	*15,0*
Corpus de la Hong-Kong City-University		
Zhikov *et al.* (sans MDL)	0,652	23,2
Zhikov *et al.* (avec leur MDL)	0,787	19,8
Notre système (sans notre étape MDL)	0,744	20,3
Notre système (avec notre étape MDL non contrainte)	0,754	**19,3**
Notre système (avec notre étape MDL contrainte)	**0,801**	19,8
Gold	*1,0*	*19,0*
Corpus de Microsoft Research		
Zhikov *et al.* (sans MDL)	0,690	37,1
Zhikov *et al.* (avec leur MDL)	0,782	31,9
Notre système (sans notre étape MDL)	0,782	33,0
Notre système (avec notre étape MDL non contrainte)	0,690	**31,1**
Notre système (avec notre étape MDL contrainte)	**0,809**	32,1
Gold	*1,0*	*30,8*
Corpus de l'Academia Sinica		
Zhikov *et al.* (sans MDL)	0,614	80,8
Zhikov *et al.* (avec leur MDL)	0,762	67,1
Notre système (sans notre étape MDL)	0,758	68,9
Notre système (avec notre étape MDL non contrainte)	0,711	**65,7**
Notre système (avec notre étape MDL contrainte)	**0,795**	67,3
Gold	*1,0*	*65,3*

Tableau 7.8 – F-score sur quatre corpus de référence mesurant la qualité des formes obtenues par les techniques proposées par Zhikov *et al.* (2010) et par les trois principales configurations de notre système. Une forme du corpus de réference est considérée comme trouvée avec succès si ses deux frontières sont correctement identifiées et qu'aucune frontière n'est proposée par le système entre les sinogrammes qui la composent.

l'une des sources d'erreurs réside dans les entités nommées, et notamment dans celles qui ont des syntaxes internes plus spécifiques. Ce n'est pas surprenant : ces syntaxes internes de ces entités n'ayant pas grand chose à voir avec la syntaxe générale, on peut s'attendre à ce qu'elles introduisent du bruit dans les modèles de segmentation généraux tout en étant difficiles à segmenter pour ces modèles. Or la structure interne de telles entités nommées (dates, adresses…) n'est pas à proprement parler du ressort de la langue en tant que système, mais plutôt de micro-systèmes paralinguistiques conventionnels [47]. Sans surprise, les corpus de référence utilisent d'ailleurs pour ces entités des conventions de segmentation souvent fort différentes. En conséquence, il est utile de disposer et de faire usage de grammaires locales telles que celles préalablement présentes dans SxPipe pour d'autres langues à séparateur typographique. L'idée en effet est d'identifier dans une phrase de pré-traitement ces entités nommés et de les segmenter en fonction de conventions qui leur sont propres, puis de faire fonctionner nos algorithmes de segmentation non supervisée sur des données au sein desquelles chacune de ces entités, dont le contenu est rendu opaque et non segmentable, peut être remplacé par une indication de son type.

Certes, de telles grammaires locales dépendent en partie de la langue [48], et peuvent être considérées comme remettant en cause le caractère non supervisé de notre approche. Ce n'est pas notre opinion, dans la mesure où le développement de ces grammaires ne concerne donc pas l'identification des frontières entre formes ressortant de la langue générale.

Nous avons ainsi étendu SxPipe pour qu'il puisse traiter un certain nombre d'entités nommées du mandarin, et réentraîné nos modèles (sans étape de MDL). Les résultats sont significativement améliorés, la f-mesure étant améliorée selon les corpus et en chiffres absolus entre 0,5% (corpus de Microsoft Research) et 4% (Corpus de l'Université de Pékin, Corpus de la Hong-Kong City-University). Nous ne donnons pas ici les résultats précis car les expériences réalisées l'ont été avec une version du système antérieure à celle ayant donné les résultats présentés dans le tableau 7.8. Mais cela confirme la pertinence qu'il y a à traiter séparément les entités typographiques par rapport au reste des données. Nous ferons le même constat à la section 8.1.4 sur l'étiquetage des entités nommées pour le français.

47. Du reste, ces conventions varient souvent d'un pays à l'autre, y compris lorsque ces pays partagent une même langue. Il n'est qu'à se rappeler les façons différentes qu'ont par exemple les québecois et les français d'écrire une adresse postale, ou les anglais et les américains d'écrire une date. Les conséquences de ce caractère conventionnel vont du reste au-delà de la seule problématique de segmentation étudiée ici. Nous y reviendrons à propos de l'étiquetage morphosyntaxique, notamment à la section 8.1.4.

48. En partie seulement, comme discuté précédemment : reconnaître correctement une URL dans un texte peut être effectué de façon identique quelle que soit la langue du texte, quand bien même certains réglages permettant d'optimiser la précision et/ou le rappel peuvent bénéficier d'adaptations dépendantes de la langue.

7.5 Éléments de conclusion

La problématique générale de l'identification des unités élémentaires d'analyse dans un flux textuel reste un champ de recherche ouvert, dont l'impact sur la qualité des traitements ultérieurs ne saurait être sous-estimé. Nous aurons l'occasion de l'illustrer à propos de l'étiquetage en parties du discours à la section 8.4 puis, pour l'analyse syntaxique, à la section 9.4. Les travaux les plus récents dans ce domaine explorent deux pistes complémentaires des travaux décrits dans ce chapitre, et que nous n'avons que partiellement explorés : les approches jointes avec des traitements aval, y compris l'analyse morphosyntaxique et syntaxique, et les approches supervisées, notamment pour la détection de mots composés. À cet égard, nous avons récemment travaillé dans le cadre de la campagne d'évaluation UD 2017 (cf. section 8.2) à l'identification supervisée, au moyen de modèles statistiques enrichis par des lexiques externes, de ce qui est appelé « tokens » et « phrases » dans les jeux de données concernés. Pour simplifier, la notion de « tokens » utilisée dans les jeux de données de cette campagne d'évaluation est la même que la nôtre pour les langues à séparateur typographique pertinent, mais correspond à la notion de forme pour les autres (chinois, japonais, vietnamien). Nous y reviendrons brièvement dans la note 51 du chapitre 8 (on pourra également se reporter à (Villemonte de La Clergerie *et al.*, 2017) pour plus de détails). Quant à la normalisation des données bruitées, tâche théoriquement indissociable de celle de l'identification des formes, elle fait également l'objet d'un nombre croissant de travaux, à la fois non-supervisés et supervisés, y compris au moyen d'approches inspirées de la traduction automatique neuronale. C'est là une direction de recherche que j'envisage d'explorer à l'avenir, y compris au sein de systèmes joints avec effectuant également l'étiquetage morphosyntaxique et l'analyse morphologique.

De façon générale, l'une des difficultés majeures des travaux sur la segmentation en phrase et l'identification et normalisation des formes réside dans la difficulté qu'il y a à definir de façon opérationnelle la notion de forme, et donc également celle de mot syntaxique, puisque nous avons choisi au début de ce chapitre de nous appuyer sur l'approximation consistant à identifier ces dernières aux formes. Les approches non-supervisées telles que celle que nous venons de décrire dans le cas du mandarin permettent de retrouver en partie la notion de forme telle que l'on peut essayer de la définir linguistiquement, mais les principes sur lesquels elles reposent (entropie de branchement, MDL) ne sont que des moyens indirects d'approcher les critères linguistiques de minimalité et d'autonomie aux niveaux morphosyntaxiques et syntaxiques évoqués au chapitre 1. Les approches supervisées et les approches par règles, de leur côté, visent à reproduire des jugements et les décisions linguistiques encodés respectivement dans les corpus d'entraînement et dans les règles utilisées. Mais de telles approches trouvent leur limite dès lors que l'on s'éloigne des types de données classiques, et en particulier, dans

le cas des approches supervisées, dès que l'on traite de corpus aux propriétés différentes de celles du corpus d'apprentissage.

Chapitre 8

Analyse morphosyntaxique et informations lexicales

Sommaire

L'étiquetage morphosyntaxique (*tagging*) est une tâche désormais classique en traitement automatique des langues, pour laquelle de nombreux systèmes ont été développés ou adaptés à un large éventail de langues. Elle consiste à associer à chaque « mot » une *étiquette morphosyntaxique* dont la granularité peut aller d'une simple catégorie morphosyntaxique, ou partie du discours, à une catégorie plus fine et enrichie par des traits morphologiques (genre, nombre, cas, temps, mode, etc.).

La section A.9 résume l'historique des recherches dans ce domaine. Il en ressort que l'utilisation d'algorithmes d'apprentissage automatique faisant usage de corpus annotés manuellement est désormais la norme pour le développement d'étiqueteurs morphologiques. Différents types d'algorithmes ont été utilisés, dont successivement les modèles de Markov cachés bigrammes puis trigrammes (Merialdo, 1994 ; Brants, 1996, 2000), les arbres de décision (Schmid, 1994 ; Magerman, 1995), les modèles de Markov à maximisation d'entropie (Maximum Entropy Markov Models, MEMM ; Ratnaparkhi, 1996) et les champs aléatoires conditionnels (Conditional Random Fields, CRF ; Constant *et al.*, 2011). Ces algorithmes d'apprentissage automatique permettent la construction de systèmes d'étiquetage pour n'importe quelle langue, pourvu que l'on dispose de données d'apprentissage adaptées [1]

Pour compléter ces données annotées, plusieurs travaux ont montré la pertinence de s'appuyer sur des informations lexicales externes, et notamment sur des lexiques morphosyntaxiques associant par exemple une partie du discours à un grand nombre de mots. Ces informations peuvent être utilisées sous forme de contraintes au moment de l'étiquetage de nouvelles données (Kim *et al.*, 1999 ; Hajič, 2000) ou pour l'extraction de traits qui complètent les traits extraits du corpus annoté dès la phase l'apprentissage (cf. par exemple Chrupała *et al.*, 2008 ; Goldberg *et al.*, 2009). C'est cette dernière approche que nous avons explorée dans nos travaux sur trois étiqueteurs morphosyntaxiques sucessifs, travaux qui font l'objet de ce chapitre.

Nous reviendrons ainsi successivement les trois étiqueteurs morphosyntaxiques que nous avons développés ces dernières années : les étiqueteurs statistiques MElt (Maximum-Entropy Lexicon-enriched Tagger, Denis et Sagot, 2009, 2010, 2012) et alVWtagger d'une part, et l'étiqueteur neuronal alNNtagger d'autre part. Ces trois étiqueteurs peuvent faire usage d'informations issues d'un lexique morphosyntaxique. Nous metterons donc

1. Certains systèmes sont largement utilisés notamment parce qu'ils sont associés à nombreux modèles faciles à télécharger et à utiliser. C'est par exemple le cas de TreeTagger (Schmid, 1994), disponible sur `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`, qui repose sur l'apprentissage d'arbres de décision, pour lequel des modèles pour plus d'une vingtaine de langues sont disponibles. Le modèle fourni pour le français constitue d'ailleurs probablement l'étiqueteur le plus utilisé dans la communauté francophone, certainement en raison de sa gratuité (sans pour autant être un logiciel libre, son code source n'étant pas disponible), de son ancienneté et de sa facilité d'utilisation, malgré des performances inférieures à plusieurs étiqueteurs plus récents comme MElt (cf. ci-dessous, ou directement Denis et Sagot (2009)).

l'accent sur l'impact des informations lexicales sur les performances de ces étiqueteurs, sur la façon d'optimiser l'utilisation de ces informations, sur le bénéfice que l'on peut en tirer pour accélérer le développement des ressources nécessaires à l'entraînement d'étiqueteurs. Nous étudierons également la portabilité de cette approche à de nombreuses langues, en tirant parti de différents jeux de données en lien avec les campagnes d'évaluation SPMRL et avec l'initiative *Universal Dependencies*, y compris les données de la campagne d'évaluation CoNLL 2017 (cf. plus bas), couvrant une cinquantaine de langues. Cette campagne a du reste mis en exergue une limite de nombre de travaux en étiquetage morphosyntaxique, y compris une partie de nos travaux sur MElt et sur alNNtagger, à savoir celle de leur évaluation. Celle-ci est en effet souvent effectuée sur des corpus déjà segmentés en « phrases » et en formes. Pourtant, dans le cas général, il ne s'agit pas là d'une configuration réaliste dès lors que l'on veut traiter de corpus bruts. Les problématiques soulevées au chapitre précédent reviennent alors au premier plan. Nous discuterons donc de ces questions, et les illustrerons, entre autres, à travers nos travaux sur l'étiquetage morphosyntaxique de corpus bruités issus du web, réalisés en particulier sur l'anglais dans le contexte de la campagne d'évaluation SANCL 2012 et sur le français dans celui du développement du *French Social Media Bank*, corpus arboré de données issues du web.

8.1 Informations lexicales et étiquetage morphosyntaxique statistique : MElt [2], [3]

MElt repose sur un modèle de Markov à maximisation d'entropie, modèle séquentiel linéaire, pour quatre raisons principales : (i) ces modèles ont de bonnes propriétés, et sont notamment faciles et raisonnablement rapides à entraîner [4] et produisent des modèles directement interprétables [5] ; (ii) les étiqueteurs reposant sur les MEMM sont parmi les plus performants sur l'anglais ; (iii) il existe des implémentations efficaces des MEMM, telles que `megam` [6] (Daumé III, 2004), utilisé par MElt ; [7] (iv) il est très simple en théorie et

2. Les premières étapes du travail présenté dans cette partie (sections 8.1.1 et 8.1.2) ont été réalisées en collaboration avec Pascal Denis. Le travail dans son ensemble a fait l'objet de plusieurs publications (Denis et Sagot, 2009, 2010, 2012 ; Sagot, 2016b,a).

3. MElt est librement disponible à l'adresse `http://lingwb.gforge.inria.fr/`. Les résultats de ce chapitre, sauf mention explicite du contraire, correspondent à la version 1.0. Le chapitre 7 présente quant à lui les résultats de travaux ultérieurs ayant conduit à la version 2.0 de MElt.

4. Contrairement, par exemple, aux champs aléatoires conditionnels (Conditional Random Fields, CRF, (Lafferty *et al.*, 2001), souvent considérés comme plus adaptés aux problèmes de prédiction séquentielle et structurée, mais plus lents à entraîner. De plus, ils obtiennent sur la tâche d'étiquetage morphosyntaxique des performances similaires aux MEMM (Constant et Tellier, 2012).

5. Contrairement, par exemple, aux étiqueteurs neuronaux, sur lesquels nous reviendrons également plus bas.

6. Librement disponible à l'adresse `http://www.cs.utah.edu/~hal/megam/`.

7. Nous verrons plus bas, cependant, que la rapidité d'entraînement peut être encore améliorée grâce au système récent Vowpal Wabbit.

possible en pratique avec `megam` d'utiliser à la place d'un modèle MEMM tout autre modèle séquentiel linéaire, et notamment un perceptron multiclasses, sans changement ni dans les données d'entrées pour l'apprentissage du modèle ni au moment de son utilisation pour l'étiquetage.

8.1.1 Modèles d'étiquetage

MElt s'appuie sur un modèle markovien à maximisation d'entropie (MEMM). Le modèle le plus simple, qui nous servira de base de comparaison, est un modèle ne faisant pas usage d'informations lexicales. Il est comparable aux systèmes de Ratnaparkhi (1996) et Toutanova et Manning (2000), à la fois quant au modèle et quant aux traits utilisés. Un avantage important de ces modèles (sur les modèles de Markov cachés, notamment) est de permettre de combiner ensemble des traits très divers, éventuellement redondants, sans qu'il soit nécessaire de faire une hypothèse d'indépendance entre eux. Notons par ailleurs qu'il s'agit d'un modèle gauche-droite, qui peut donc exploiter les étiquettes déjà attribuées aux mots dans le contexte gauche pour étiqueter le mot courant.

L'objectif de notre travail est d'étudier l'impact de l'utilisation d'un lexique externe. En effet, mis à part des traits sur la forme des mots (préfixes et suffixes [8]), seul un lexique externe peut fournir des informations sur le contexte droit (par exemple, la ou les catégories connues par le lexique pour le mot qui suit le mot courant). De plus, un mot inconnu du corpus pouvant être connu du lexique externe, il y a là une source d'amélioration pour l'étiquetage de cette catégorie de mots et les mots qui sont dans leur voisinage. MElt s'appuie donc sur un modèle qui étend le modèle de base par l'utilisation de traits supplémentaires qui reposent sur un lexique externe, et notamment des traits codant la ou les catégories connues par le lexique pour le mot courant, le mot qui le suit, le mot qui le précède, des combinaisons de ces catégories, et d'autres. Nous renvoyons à (Denis et Sagot, 2012) pour une description détaillée des classes de traits utilisés.

8.1.2 Premières expériences sur le français

L'une des motivations pour le développement de MElt était le constat que l'étiquetage morphosyntaxique avait longtemps été moins étudié pour le français que pour d'autres langues d'importance équivalente. C'est la raison pour laquelle nos premières expériences avec MElt ont été réalisées sur le français.

Nous nous sommes appuyés pour ces premières expériences sur le Corpus Arboré de Paris 7 (ou French TreeBank, ci-après FTB ; Abeillé *et al.*, 2003), dans sa variante dite FTB-UC (Candito et Crabbé, 2009). C'est donc un corpus déjà segmenté en unités de traitements (« phrases ») et en unités élémentaires (« mots »). À cet égard, il diffère du FTB originel en

8. Ces termes sont employés ici en un sens non linguistique : ce sont seulement les séquences formées des *n* premiers ou derniers caractères du mot.

ceci que tous les composés qui ne correspondent pas à une séquence régulière de parties du discours sont fusionnés en une unitée unique, alors que les autres sont représentés par des séquences de plusieurs unités. Par abus de langage, nous dénoterons par le terme *mot* les unités ainsi obtenues, évitant ainsi d'avoir à nous prononcer sur la nature linguistique, si tant est qu'il en existe une qui soit cohérente et homogène, de ces unités. Insistons à nouveau sur le fait que partir d'une telle segmentation en mots (au sens que nous venons de définir) n'est pas une situation très vraisemblable dans un contexte d'application réelle, comme discuté au chapitre précédent. La version du FTB utilisé dans cette section contient 350 931 mots pour 12 351 phrases et s'appuie sur un jeu de 29 parties du discours, avec une granularité intermédiaire entre catégories et sous-catégories de la version d'origine du FTB [9]. Nous utilisons le découpage en sous-corpus utilisé depuis Crabbé et Candito (2008) : entraînement (80%), développement (10%) et test (10%), que nous appellerons désormais respectivement FTB-TRAIN, FTB-DEV et FTB-TEST. Cette division du FTB en trois sous-corpus sera utilisée dans ce chapitre et dans le suivant, tant pour des tâches d'étiquetage morphosyntaxique que pour des tâches d'analyse syntaxique. Les tailles respectives de ces sous-corpus sont détaillées à la table 8.1, ainsi que les nombres et proportions de mots inconnus du lexique externe utilisé, le Le*fff* [10].

Section	**Nb. de phrases**	**Nb. de tokens**	**Nb. de tokens inconnus**	**Nb. de tokens inconnus et absents du Le*fff***
FTB-TRAIN	9 881	278 083		
FTB-DEV	1 235	36 508	1 790 (4,9%)	604 (1,7%)
FTB-TEST	1 235	36 340	1 701 (4,7%)	588 (1,6%)

TABLEAU 8.1 – Jeux de données

La source d'informations lexicales que nous avons utilisée est en effet le Le*fff* (cf. section 2.1). Nous en avons extrait 502 223 entrées distinctes de la forme *(forme, étiquette)*, les étiquettes correspondant après conversion au jeu de 29 étiquettes de la variante du FTB décrite ci-dessus et utilisée pour l'apprentissage.

8.1.2.1 Évaluation comparative

Nous avons comparé les résultats de $\text{MElt}_{\text{fr}}^{\text{FTB-UC,NoLex}}$ (modèle n'utilisant pas de lexique externe), de $\text{MElt}_{\text{fr}}^{\text{FTB-UC,C}}$ (modèle avec lexique externe utilisé comme source de contraintes au décodage) et de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ (modèle avec lexique externe utilisé comme sources de traits) à divers autres étiqueteurs disponibles à l'époque de ces expériences

9. Ces 29 étiquettes améliorent les catégories principales par des informations sur le mode des verbes, ainsi que par quelques traits lexicaux supplémentaires. Ce jeu d'étiquettes est celui qui conduit aux meilleurs résultats d'analyse syntaxique probabiliste pour le français (Crabbé et Candito, 2008 ; ce jeu d'étiquettes y est nommé TREEBANK+).

10. Il s'agit des mots inconnus à la fois sous leur forme d'origine *et* sous leur forme minusculisée.

(2012). Les deux premiers d'entre eux n'utilisent pas le Le*fff*. Ils ont tous été (ré)entraînés sur FTB-TRAIN [11] :

- UNIGRAM, un étiqueteur *baseline* qui assigne aux mots présents dans le corpus d'entraînement l'étiquette la plus fréquemment trouvée dans le corpus ; pour les autres mots, il utilise l'étiquette la plus fréquente du corpus (ici, *NC*) ;
- TreeTagger, un étiqueteur statistique qui repose sur les arbres de décision (Schmid, 1994) ;
- UNIGRAM$_{\text{Le}fff}$, comme UNIGRAM, est un modèle unigramme qui repose sur le corpus d'apprentissage, mais qui utilise le Le*fff* pour étiqueter les mots inconnus : parmi les étiquettes que le Le*fff* associe à un mot inconnu du corpus, l'étiquette la plus fréquente à l'échelle de tout le corpus est utilisée ; les mots qui sont inconnus et du corpus et du Le*fff* reçoivent l'étiquette la plus fréquente (ici, *NC*) ;
- TreeTagger$_{\text{Le}fff}$ est une variante de TreeTagger où le Le*fff* est fourni comme lexique externe ;
- LGtagger (Constant *et al.*, 2011 ; Constant et Sigogne, 2011) est un étiqueteur reposant sur un modèle CRF et sur plusieurs lexiques externes, dont le Le*fff* [12] ;
- F-BKY, une instance de l'analyseur syntaxique de Berkeley tel qu'adapté au français par Crabbé et Candito (2008), et utilisée ici comme étiqueteur (cf. discussion dans l'introduction de ce chapitre).

Les résultats de cette comparaison sur le corpus de test du FTB font l'objet du tableau 8.2.

Étiqueteur	**Précision globale (%)**	**Précision sur les mots inconnus (%)**
UNIGRAM	91,90	24,50
TreeTagger	96,14	75,77
UNIGRAM$_{\text{Le}fff}$	93,40	55,00
TreeTagger$_{\text{Le}fff}$	96,55	82,14
LGtagger	97,7	—
F-BKY	97,25	82,90
MElt$_{\text{fr}}^{\text{FTB-UC,NoLex}}$	97,00	86,10
MElt$_{\text{fr}}^{\text{FTB-UC,C}}$	97,25	86,47
MElt$_{\text{fr}}^{\text{FTB-UC}}$	**97,75**	**91,36**
MElt$_{\text{fr}}^{\text{FTB-UC}}$ version 2016 [13]	**97,87**	**91,66**

TABLEAU 8.2 – Comparaison des performances de divers étiqueteurs en partie du discours pour le français (évaluation sur FTB-TEST)

11. Le décodage est fait avec un *beam* de 3, des expériences préliminaires sur le corpus FTB-DEV n'ayant pas montré de changement significatif avec des valeurs plus élevées.

12. Nous n'avons pas réentraîné cet étiqueteur. Les résultats présentés à la table 8.2 sont ceux donnés par Constant *et al.* (2011) et Constant et Sigogne (2011).

13. Cf. section 8.1.3 et note 27.

Parmi les étiqueteurs ne faisant pas usage du Le*fff*, on constate que $\text{MElt}_{\text{fr}}^{\text{FTB-UC,NoLex}}$ atteint déjà une précision [14] de 97,00%, avec 86,10% sur les mots inconnus. Ceci est significativement meilleur que TreeTagger, avec un gain de plus de 10% sur les mots inconnus [15]. Parmi les étiqueteurs faisant usage du Le*fff*, le meilleur d'entre eux est $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$, avec une précision de 97,75% globalement et de 91,36% sur les mots inconnus. Ces deux résultats constituent des améliorations significatives de 0,75% (25% en relatif) et 5,26% (38% en relatif) par rapport au modèle sans Le*fff*. Ces scores sont meilleurs que ceux de tous les étiqueteurs que nous avons pu tester à cette époque (2008–2012), y compris l'analyseur F-BKY qui exploite des résultats d'analyse syntaxique probabiliste, et ce, avec un écart significatif [16]. Le fait que $\text{MElt}_{\text{fr}}^{\text{FTB-UC,C}}$ conduise à des résultats inférieurs à $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ peut probablement s'expliquer par les deux différences suivantes : d'une part $\text{MElt}_{\text{fr}}^{\text{FTB-UC,C}}$ ne bénéficie pas des informations supplémentaires dont dispose $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ à propos du contexte droit du mot courant, et d'autre part $\text{MElt}_{\text{fr}}^{\text{FTB-UC,C}}$ est contraint de respecter les indications du lexique, là où $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ peut les remettre en cause puisque ces indications sont utilisées sous forme de traits.

8.1.2.2 Analyse d'erreurs

Afin d'étudier d'où viennent les erreurs de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ et de comprendre comment ce score de 97,75% pourrait être amélioré, nous avons décidé d'examiner à la main ses 200 premières erreurs sur FTB-DEV puis de les classer. La typologie des erreurs que nous avons retenue et la répartition de ces dernières selon cette typologie est donnée à la table 8.3.

Les résultats montrent que ce score de 97,75% peut être amélioré. Tout d'abord, des techniques classiques de reconnaissance des entités nommées pourraient contribuer à diminuer le nombre d'erreurs liées à ces entités, lesquelles représentent plus d'un quart du total des erreurs. Nous reviendrons à la section 8.1.4 sur des expériences que nous avons menées dans cette direction. De plus, des motifs simples pourraient permettre de remplacer tous les nombres par un ou plusieurs marqueurs génériques, à la fois dans les données d'apprentissage et dans les données d'évaluation. En effet, conserver les

14. Le terme d'exactitude serait ici plus approprié, chaque mot recevant exactement une étiquette quoi qu'il arrive. Pour simplifier la lecture, et par abus de langage, nous utiliserons toutefois le terme de précision à cette fin.

15. On peut avancer plusieurs hypothèses pour expliquer des écarts si importants sur l'étiquetage des mots inconnus. Tout d'abord, l'estimation des paramètres dans un modèle à maximisation d'entropie est moins sujette au problème du manque de données pour certains traits ou certaines valeurs de traits que d'autres approches comme les arbres de décision (utilisés par TreeTagger). Par ailleurs, TreeTagger n'est pas en mesure de faire autant de généralisations que $\text{MElt}_{\text{fr}}^{\text{FTB-UC,NoLex}}$ sur les traits internes, puisqu'il ne prend en compte que les suffixes, et ce, uniquement sur les mots inconnus.

16. Une adaptation au français de Morfette (Chrupała *et al.*, 2008) utilisant le FTB et le Le*fff* a été réalisée par G. Chrupała et D. Seddah (c.p.). Leur précision est comparable à celle de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ (sur les mêmes jeux de données, Henestroza et Candito (c.p.) ont obtenu 97,68%). Sur d'autres variantes du FTB (tokenisation d'origine), Chrupała et Seddah (c.p.) obtiennent 97,9%. Ces comparaisons sont à nuancer dans la mesure où des informations supplémentaires (les lemmes) sont extraites des corpus d'apprentissage et prises en compte dans ce modèle.

Type d'erreur		Fréquence
Erreurs standard	Adjectif vs. participe passé	5,5%
	Erreurs sur *de*, *du*, *des*	4,0%
	Autres erreurs	34,0%
Erreurs sur les nombres		15,5%
Erreurs liées aux entités nommées		27,5%
Faux négatifs	Erreurs d'étiquteage dans le FTB-DEV	8,5%
	Erreur de tokenisation dans le FTB-DEV (mot tronqué)	0,5%
	Les étiquettes du FTB-DEV et de MElt semblent toutes deux valides	4,5%

TABLEAU 8.3 – Analyse manuelle des 200 premières erreurs de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ sur le corpus de développement FTB-DEV

nombres tels quels ne peut que conduire à des problèmes de dispersion des données, ce qui empêche l'algorithme d'apprentissage de modéliser correctement cette tâche complexe qu'est l'étiquetage des nombres (dans le FTB-UC ils peuvent être des déterminants, des noms, des adjectifs ou des pronoms). Par ailleurs, certaines classes d'erreurs, comme celles sur les mots *de*, *du* et *des*, pourraient être traitées en partie par des motifs dépendants de la langue, à l'image des travaux de Urieli (2014) sur l'étiquetage de *que* ou, quoi qu'avec une granularité plus fine que celle du jeu d'étiquettes utilisé ici, ceux de Danlos (2005) sur *il* (pronom impersonnel ou personnel). Enfin, pas moins de 13,5% des erreurs identifiées sont en fait liées aux annotations du corpus, soit à cause d'erreurs dans le FTB-DEV (9%), soit à cause de cas peu clairs, où l'étiquette présente dans le FTB-DEV et celle proposée par MElt semblent toutes deux convenir.

Nous avons également mené des analyses complémentaires pour étudier l'impact des différents types de traits lexicaux et l'impact de la granularité du jeu d'étiquettes. Nous renvoyons pour cela respectivement à (Denis et Sagot, 2012) et à (Denis et Sagot, 2010). Nous avons également mené quelques expériences sur d'autres langues (anglais et espagnol) avec cette même version de MElt, pour lesquelles nous renvoyons à (Denis et Sagot, 2012).

8.1.2.3 Impact relatif de la taille du corpus d'entraînement et de celle du lexique externe

Les résultats obtenus par $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ sur le français bénéficient largement de la taille raisonnablement grande de son corpus d'entraînement et de la bonne couverture du lexique externe. Cependant, de telles ressources ne sont pas toujours disponibles pour une langue donnée, et notamment pour les langues dites peu dotées. De plus, l'amélioration significative que l'on observe lorsque l'on rajoute les traits issus du lexique externe montre que les informations que contient un tel lexique sont pertinentes pour la tâche

d'étiquetage morphosyntaxique. La question se pose alors de savoir si ces informations lexicales peuvent compenser le handicap que constitue un petit corpus d'entraînement. À l'inverse, il serait intéressant de savoir quel est l'impact de la taille du lexique externe sur la qualité des résultats d'étiquetage.

Nous avons donc procédé à une série d'expériences consistant à entraîner $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ sur différents sous-corpus d'entraînement et avec différents sous-lexiques. L'extraction des sous-corpus à partir de son corpus d'entraînement standard, FTB-TRAIN, s'est faite de façon simple, en choisissant pour corpus de taille s, mesurée en nombre de phrases, les s premières phrases du corpus complet. Mais l'extraction de sous-lexiques à partir du Le*fff* est moins évidente. Nous avons décidé d'extraire des sous-lexiques de taille croissante d'une manière qui reproduit approximativement le processus de développement d'un lexique morphosyntaxique. Pour cela, nous avons utilisé $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ pour étiqueter un gros corpus, le corpus de l'Est Républicain, librement disponible [17]. Nous avons lemmatisé ce corpus à l'aide du lemmatiseur fourni avec MElt. Enfin, nous avons trié les couples *(lemme, étiquette)* par fréquence décroissante. Nous avons alors décidé qu'un lexique de taille l, mesuré en nombre de lemmes, serait constitué de l'ensemble des formes fléchies des l couples *(lemme, étiquette)* les plus fréquents. L'idée sous-jacente est qu'il est peu réaliste que le développement d'un lexique morphosyntaxique se fasse forme fléchie par forme fléchie, et qu'au contraire on s'attend, y compris dans une architecture comme Alexina, à ce que le développeur de lexique rajoute progressivement des entrées lexicales de niveau lemme, qui donnent naissance, d'une façon ou d'une autre, à l'ensemble de ses formes fléchies.

Nous avons ré-entraîné différentes variantes de MElt avec 8 tailles différentes de corpus d'entraînement et 9 lexiques externes de tailles différentes, y compris le lexique vide [18] (toutes les combinaisons entre taille du lexique et taille du corpus ont été testées) [19]. Pour chacun des 72 étiqueteurs ainsi obtenus nous avons conservé sa précision globale et sa précision sur les mots inconnus telles qu'évaluées sur FTB-TEST.

Avant de comparer l'impact respectif du développement manuel du lexique externe et du corpus pour optimiser les performances de l'étiqueteur, il nous faut estimer

17. `http://www.cnrtl.fr/corpus/estrepublicain/`

18. Nous avons réalisé des expériences sans lexique externe, mais pas sans corpus d'entraînement. En effet, l'annotation morphosyntaxique avec un lexique morphologique mais sans corpus d'entraînement est une tâche très différente, identifiée depuis les années 1970 et traitée de plus en plus régulièrement (Merialdo, 1994 ; Smith et Eisner, 2005 ; Ravi et Knight, 2009). À cet égard, MElt a été utilisé dans une expérience simple sur le kurde kurmanji, une langue iranienne peu dotée (Walther *et al.*, 2010). Nous avons projeté le lexique Alexina du kurde kurmanji, KurLex, de couverture moyenne, sur un corpus brut. Nous avons ensuite désambiguïsé l'annotation ambiguë ainsi obtenue de trois façons différentes, puis fusionné ces annotations pour produire un corpus annoté (bruité), et entraîné MElt sur ce corpus et avec KurLex comme lexique externe. Malgré la simplicité de nos trois méthodes de désambiguïsation des annotations, nous avons obtenu une précision de 85,7% avec un jeu de 36 étiquettes.

19. Tailles des lexiques (en lemmes) : 0, 500, 1 000, 2 000, 5 000, 10 000, 20 000, 50 000, 110 000 (Le*fff* complet). Tailles des corpus : 50, 100, 200, 500, 1 000, 2 000, 5 000, 9 881 (FTB-TRAIN complet).

quantitativement les coûts de ce développement manuel. Dans (Marcus *et al.*, 1993), les auteurs font état de vitesses d'annotation manuelle en parties du discours qui est au-delà de 3 000 mots par heure au cours du développement du Penn TreeBank. Cette vitesse est atteinte après une période d'un mois (pour 15 heures d'annotation par semaine, soit environ 60 heures), période au cours de laquelle la qualité de l'étiqueteur utilisé comme pré-annotateur s'améliore progressivement. Nous avons nous-même estimé le rythme de la validation manuelle de corpus pré-annotés, au cours d'expériences visant à étudier l'impact de la qualité de la préannotation sur la vitesse et la qualité de la validation manuelle pour le développement de corpus étiquetés morphosyntaxiquement (Fort et Sagot, 2010). Nous avons obtenu, une fois la phase de rôdage terminée, des vitesses d'un peu moins de 3 minutes pour un bloc de 10 phrases soit environ 6 000 mots à la minute. C'est le double du chiffre précédemment mentionné, mais cela ne concerne que des efforts très brefs, de l'ordre de 3 minutes. C'est une chose que d'annoter à une certaine vitesse pendant 3 minutes, et une autre de tenir cette vitesse pendant une heure entière. Il est donc raisonnable d'en rester aux chiffres plus globaux donnés par Marcus *et al.* (1993) et de faire l'hypothèse d'une vitesse de développement qui est de l'ordre de 1 000 mots (30 phrases) par heure au départ, puis qui accélère jusqu'à 3 000 mots (100 phrases) par heure après que 5 000 phrases ont été annotées.

Pour le développement du lexique externe, des techniques telles que celles décrites au chapitre 3, notamment à la section 3.1.1, permettent de valider rapidement des entrées lexicales proposées automatiquement. Le travail manuel est alors limité à cette étape de validation, qui prend au plus 2 à 3 secondes par lemme, soit 1 500 lemmes par heure.

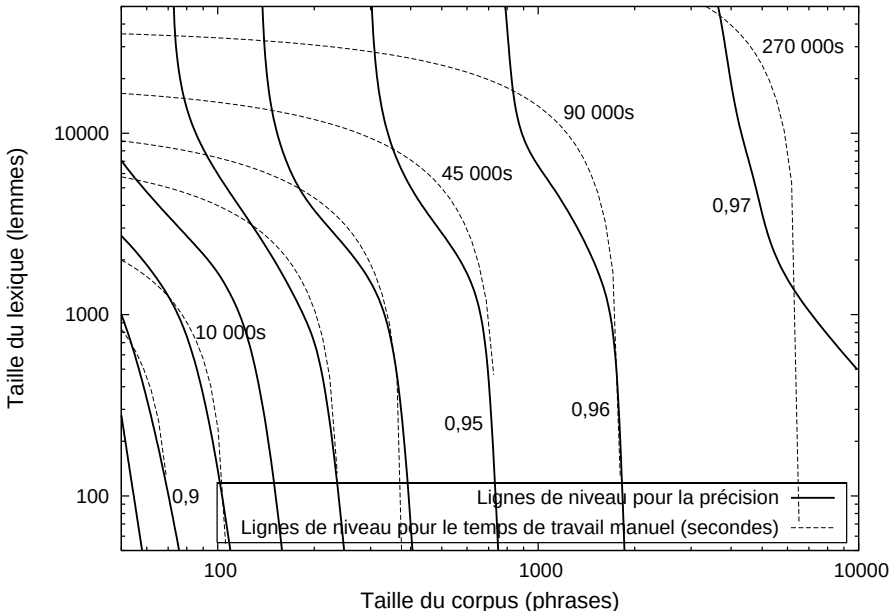
La figure 8.1 compare les courbes de niveaux [20] pour deux fonctions de la taille du corpus d'apprentissage et de celle du lexique externe : la précision de l'étiqueteur et le temps de développement manuel des ressources [21]. On peut faire de ces graphes les commentaires suivants :

— pendant les premières phrases de développement (moins de 3 heures de travail manuel), la distribution de ce travail entre développement de lexique et annotation de corpus n'a pas d'impact significatif sur la précision globale de l'étiquetage, mais la précision sur les mots inconnus est meilleure si l'on passe autant ou plus de temps sur le lexique que sur le corpus ;

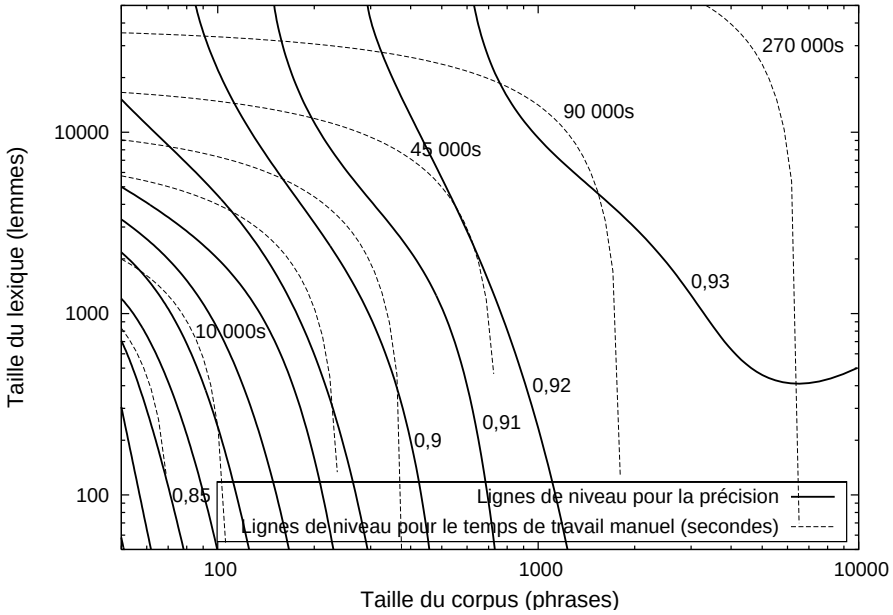
— dans les phases ultérieures du développement, l'approche optimale consiste à développer à la fois le lexique et le corpus ; ceci améliore à la fois la précision globale et la précision sur les mots inconnus ; toutefois, si l'on doit se contenter

20. Obtenues au moyen du mode `bspline` de l'algorithme de calcul de courbes de niveau de l'outil `gnuplot`.

21. Les temps de développement par phrase annotée et par entrée lexicale mentionée précédemment conduisent à la formule suivante, dans laquelle le temps de développement $t(s,l)$ (en secondes) dépend de la taille s du corpus annoté et du nombre l de lemmes présents dans le lexique : $t(s,l) = 36s + 8400 \cdot log(s/100 + 1) + 2.4 \cdot l$.



(a) Tous les mots



(b) Mots inconnus seulement

FIGURE 8.1 – Courbes de niveau pour deux fonctions de la taille du corpus d'apprentissage et de celle du lexique externe : précision de l'étiqueteur et temps de développement manuel des ressources. Dans (a), la précision de l'étiqueteur est sa précision globale. Dans (b), il s'agit de la précision sur les seuls mots inconnus.

de développer seulement l'une de ces deux ressources, il vaut mieux, de loin, développer un corpus ;

— utiliser un lexique morphosyntaxique externe améliore la qualité de l'étiquetage sur les mots inconnus quelle que soit l'étape de développement ;

— pour atteindre un niveau de performance donné, disposer d'un lexique externe diminue la taille du corpus d'entraînement nécessaire d'un facteur 2, parfois 3.

Ces résultats montrent pour MElt — mais nous pensons que la portée est plus générale — la pertinence du développement et de l'utilisation d'un lexique morphosyntaxique pour améliorer la qualité de l'étiquetage morphosyntaxique, et ce depuis les premières phases du développement des ressources jusqu'à l'optimisation à long terme de l'étiqueteur.

8.1.3 Expériences multilingues

Dans une nouvelle série d'expériences (Sagot, 2016b,a), nous avons cherché à évaluer la portabilité de MElt à d'autres langues, et notamment étudier la systématicité de l'amélioration induite par l'utilisation d'un lexique externe. Nous avons donc cherché des ensembles de corpus pour de nombreuses langues facilement disponibles et exploitables. Or deux ensembles de corpus annotés syntaxiquement ont été développés récemment, en grande partie mais pas seulement à partir de corpus pré-existants : les corpus SPMRL [22] (Seddah *et al.*, 2013b) et les corpus du projet *Universal Dependencies* [23] (désormais UD), disponibles à l'époque de ces expériences dans leur version 1.2. De chacun de ces corpus peuvent immédiatement être dérivés des corpus annotés morphosyntaxiquement. Nous avons donc mené nos expériences sur ces corpus.

Les données SPMRL couvrent 9 langues : arabe, basque, français, allemand, hébreu, hongrois, coréen, polonais et suédois. De son côté, le projet UD donne accès aujourd'hui à 37 corpus annotés syntaxiquement couvrant 33 langues distinctes. Contrairement aux données de la campagne SPRML, le projet UD cherche à harmoniser les annotations d'une langue à l'autre. Au niveau morphosyntaxique, cela se traduit par l'utilisation d'un jeu d'étiquette commun, les 17 *Universal Parts-Of-Speech* [24]. La pertinence linguistique d'un tel jeu d'étiquettes conçu pour être appliqué à toutes les langues est certainement discutable, mais l'initiative UD n'en est pas moins utile en pratique.

Parmi les données SPMRL, celles concernant le français, l'allemand, le suédois, le hongrois, le basque et le coréen [25] ont pour feuilles des tokens, unités typographiques simples à reproduire à partir du texte brut et permettant ainsi l'apprentissage de modèles d'étiquetage plus réalises — nous reviendrons sur cette question à la section 8.4. Le corpus

22. `http://www.spmrl.org`.

23. `http://universaldependencies.org`.

24. Cf. `http://universaldependencies.org/u/pos/all.html`, raffinement de (Petrov *et al.*, 2012).

25. Les données de l'arabe et de l'hébreu ont été segmentées de façon non triviale, les feuilles des arbres ne sont donc pas des tokens à proprement parler.

SPMRL du français, dérivé du FTB et que nous appellerons FTB-SPMRL, contient 5 000 phrases de plus pour l'entraînement, dispose de sous-corpus de développement et de test deux fois plus volumineux que dans le FTB, et inclut une annotation en mots syntaxiques multi-tokens (que nous n'avons pas encore, à ce jour, exploitée dans le cadre de nos travaux).

Nous avons donc entraîné MElt sur tous les corpus SPMRL et UD (version 1.2) [26], ainsi que sur deux corpus de référence qui ne font pas partie des deux ensemble précédents : le Penn TreeBank pour l'anglais et le Prague Dependency Treebank (version 3.0) pour le tchèque.

Ces expériences ont bénéficié d'améliorations dans MElt. Tout d'abord, nous avons intégré de nouvelles classes de traits (préfixes et suffixes du mot à droite du mot courant). Ensuite, nous avons fait varier certaines propriétés de l'espace des traits utilisés au cours d'expériences menées sur les sections d'entraînement et de développement du FTB-SPMRL. Nous avons notamment fait varier la longueur maximale des préfixes et des suffixes pris en compte par le modèle (pour le mot courant et celui à sa droite). Nous avons ensuite choisi comme nouveaux paramètres par défaut dans MElt des paramètres qui donnaient de bons résultats sur le sous-corpus de développement de FTB-SPMRL. Les expériences ci-dessous utilisent toutes ce nouveau jeu de traits [27].

Afin de comparer nos résultats à un étiqueteur plus récent et de niveau état-de-l'art, nous avons entraîné sur les mêmes corpus l'étiqueteur Marmot, système reposant sur les CRF et qui pouvait être considéré alors (fin 2015) comme le meilleur système d'étiquetage disponible (Müller *et al.*, 2013 ; Müller et Schütze, 2015) [28].

Nous avons fourni à MElt et à Marmot, dès lors que cela était possible, un lexique externe comme source d'informations complémentaires. Nous ne détaillerons pas ici l'origine de chacun de ces lexiques (cf. Sagot, 2016a et, pour le description des travaux ultérieurs sur ces lexiques, Sagot, 2018b). Un certain nombre de ces lexiques font partie de notre collection de lexiques Alexina, et ont été mentionnés au chapitre 3. D'autres, librement disponibles, ont été téléchargés et convertis automatiquement dans ce même

26. À quelques exceptions près. En effet, les traits utilisés par ces systèmes s'appuient en partie sur la granularité fines des systèmes d'écriture alphabétiques, qui permet d'extraire des préfixes et des suffixes pertinents. Nous n'avons donc pas pris en compte dans nos expériences les données coréennes et japonaises, dont les systèmes d'écriture ne sont pas alphabétiques.

27. Ces améliorations sur les traits utilisés ont également conduit à une amélioration des scores de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$, atteignant 97,88% de précision, résultat qui est aujourd'hui, à notre connaissance, état-de-l'art pour le FTB-UC (97,87% sans le *wrappeur* de gestion des entités nommées décrit à la section 8.1.4, ce qui est une différence non-significative ; ce score est indiqué dans le tableau 8.2 sous l'intitulé « $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ version 2016 »).

28. Nous aurions pu également réaliser des comparaisons avec d'autres étiqueteurs libres, spécifiques ou non au français, tels que LIA_tagg (Nasr *et al.*, 2004), le Stanford Tagger (Toutanova et Manning, 2000 ; Manning, 2011), LGtagger (Constant *et al.*, 2011 ; Constant et Sigogne, 2011) ou Morfette (Chrupała *et al.*, 2008). Pour plusieurs de ces étiqueteurs, un modèle d'étiquetage pour le français peut être téléchargé. Il s'appuie souvent en tout ou partie sur les informations lexicales du L*efff*.

formalisme, avant extraction de lexiques morphosyntaxiques (couples forme–catégorie) exploitables par MElt et Marmot (cf. section 3.1.2).

Les résultats de ces expériences sont fournis à la figure 8.4. Tout d'abord, MElt est, en moyenne, moins performant que Marmot lorsqu'ils sont entraînés uniquement sur les corpus d'entraînement, sans lexique externe. Une légère corrélation peut être observée entre d'une part l'écart entre MElt et Marmot et d'autre part la taille du corpus d'entraînement, MElt ayant tendance à être meilleur sur de petits corpus d'entraînement. Une fois prises en compte les informations lexicales externes, la situation s'inverse. L'amélioration est significative pour les deux systèmes par rapport aux modèles n'en faisant pas usage, mais cette amélioration est bien plus élevée pour MElt, au point que ce dernier passe en moyenne devant Marmot.

Il est délicat de tirer des conclusions générales à partir de tels résultats. Il nous semble toutefois que la situation ne doit pas être très différente de l'affirmation suivante : le modèle statistique de type CRF sous-jacent à Marmot est plus performant que le MEMM sur lequel s'appuie MElt ; toutefois, la façon dont les informations lexicales externes peuvent être intégrées aux modèles MElt est suffisamment performante pour permettre souvent à ce dernier de passer devant Marmot en présence de telles ressources. Dans ces cas-là, le modèle produit par MElt constitue l'état de l'art à la date de ces expériences (début 2016).

8.1.4 Gestion des entités nommées

Comme nous l'avons déjà noté à la section 7.4.3, il n'est pas clair qu'un modèle de langue, et notamment un modèle d'étiquetage morphosyntaxique (qu'il repose d'ailleurs sur des règles ou sur un modèle statistique), soit adapté au contenu des mentions d'entités, dans la mesure où ces dernières ont une structure interne qui ne relève pas de la langue générale. En conséquence, ce que nous avions dit sur la segmentation du mandarin s'applique d'ailleurs ici à l'étiquetage morphosyntaxique : l'annotation des tokens constituant les mentions d'entités est souvent délicate, les critères linguistiques guidant l'annotation des énoncés s'appliquant mal au contenu des mentions. En réalité, l'inventaire même d'étiquettes est parfois inadapté : dans une date comme `samedi 12 septembre 2015`, savoir si `12` et `2015` sont à étiqueter comme des noms ou des adjectifs relève de la convention *ad hoc* plus que de décisions linguistiquement fondées. Il en résulte du reste une grande incohérence dans de nombreux corpus annotés manuellement, y compris dans le FTB. Un modèle statistique tel que MElt ne peut qu'éprouver des difficultés à apprendre comment annoter de telles entités à partir d'annotations de référence qui sont incohérentes entre elles, et ce d'autant plus que rien ne permet à MElt de détecter explicitement qu'il est en présence d'une mention d'entité de tel ou tel type. Une autre conséquence de cela est que MElt ne peut pas apprendre

Jeu de données	Langue	nb. mots d'entr.	nb. d'ét.	Lexique	MElt		MElt vs. Marmot	
					Sans lex.	Av. lex.	Sans lex.	Av. lex.
Corpus du projet Universal Dependencies (version 1.2)								
ar	Arabe	225 853	16	—	98,39	—	−0,29*	—
bg	Bulgare	124 474	16	Multext–East	97,75	98,15	**+0,11**	**+0,10**
cs	Tchèque	1 175 374	18	Morfflex (extr.)	98,01	98,58	−0,32*	**+0,10***
cu	Vieux–slave	46 025	13	—	96,24	—	−0,02	—
da	Danois	88 979	17	STO	95,48	96,30	−0,08	**+0,14**
de	Allemand$_{web}$	274 345	16	DeLex	92,74	93,43	−0,11	**+0,33***
el	Grec moderne	47 449	11	DELA_gr	97,65	98,08	**0,00**	**+0,09**
en	Anglais$_{web}$	204 586	17	EnLex	94,06	94,60	−0,31*	**+0,05**
es	Espagnol$_{web}$	389 703	17	Le*ff*e	95,32	95,57	**+0,18**	**+0,33**
et	Estonien	7 687	15	Multext–East	89,64	94,46	**+0,52**	**0,00**
eu	Basque	72 974	16	—	94,72	—	**+0,05**	—
fa	Persan	122 093	16	PerLex	96,72	97,17	**+0,29***	**+0,20**
fi	Finlandais	162 721	15	—	93,24	—	−2,10*	—
fi$_{ftb}$	Finlandais	127 980	15	—	93,29	—	−1,24*	—
fr	Français$_{web}$	366 138	18	Le*fff*	95,81	96,14	−0,32	−0,20
ga	Irlandais	16 701	16	inmdb	92,38	92,75	**+0,99***	**+1,15***
got	Gothique	44 722	13	—	95,48	—	−0,22	—
grc	Grec class.	196 083	13	Diogenes	93,64	94,03	−0,34*	−0,29*
grc$_{proiel}$	Grec class.	166 061	13	Diogenes	96,74	97,21	−0,33*	−0,01
he	Hébreu	167 176	17	—	95,85	—	**+0,05**	—
hi	Hindi	281 057	16	—	96,29	—	−0,03	—
hr	Croate	78 817	14	HML	95,08	96,70	−0,07	**+0,51***
hu	Hongrois	20 764	16	Multext–East	94,42	94,86	**+0,07**	−0,04
id	Indonésien$_{web}$	97 531	16	Kateglo	93,74	93,83	**+0,11**	**+0,01**
it	Italien	265 992	18	Morph_it	97,44	97,82	−0,35*	−0,21*
la	Latin	37 819	12	Diogenes	93,61	94,70	**+0,63***	**+1,16***
la$_{proiel}$	Latin	132 376	13	Diogenes	96,68	96,83	−0,07	−0,16
la$_{ittb}$	Latin	246 573	14	Diogenes	98,84	99,04	−0,16	**+0,02**
nl	Néerlandais	188 882	16	Alpino	90,17	90,51	**+0,59**	**+0,36**
no	Norvégien	244 776	17	OrdBank	96,68	97,58	−0,58*	−0,04
pl	Polonais	69 499	13	PolLex	96,12	97,77	−0,09	**+0,30**
pt	Portugais	201 845	17	Labellex_pt	97,38	97,56	−0,05	**+0,17**
ro	Roumain	9 291	17	Multext–East	91,09	94,35	**+2,02***	**+2,03***
sl	Slovène	112 334	16	SloLeks	96,05	97,53	−0,18	**+0,30***
sv	Suédois	66 645	15	Saldo	95,97	96,90	−0,06	**+0,10**
ta	Tamoul	6 329	14	—	89,14	—	**+1,51***	—
Corpus de la campagne SPMRL								
TIGER	Allemand	77 220	54	DeLex	96,93	97,19	−0,43*	−0,34*
BDT	Basque	25 136	46	—	95,77	—	−0,10	—
FTB-spmrl	Français	443 113	33	Le*fff*	97,12	97,36	-0,17*	-0,05
HCTB	Hébreu	15 971	50	—	93,79	—	−0,30*	—
SzTB	Hongrois	40 782	23	Multext–East	96,68	96,94	−0,30*	**+0,01**
Składnica	Polonais	21 793	29	PolLex	96,44	97,55	**+0,02**	**+0,35***
Talbanken	Suédois	76 332	25	Saldo	96,60	97,54	**+0,23**	**+0,42***
PTB	Anglais	43 210	45	EnLex	96,81	97,01	−0,43*	−0,30*
PDT3.0	Tchèque	364 636	59	Morfflex (extr.)	98,65	99,29	−0,04	**+0,27***

Tableau 8.4 – Exactitude de MElt et de Marmot (en %) sur divers corpus. Les deux dernières colonnes donnent l'écart entre MElt et de Marmot. Un écart positif indique que MElt est meilleur, et est mis en évidence typographiquement ; un écart significatif ($p < 0{,}05$) est suivi d'un astérisque.

directement les généralisations pertinentes concernant le contexte d'apparition de tel ou tel type d'entité.

Nous avons donc mené sur le FTB-SPMRL des expériences consistant à découpler l'étiquetage de ce qui relève de la langue générale de l'étiquetage de l'intérieur de certaines mentions d'entités (adresses e-mail, URL, dates, horaires, adresses, expressions monétaires, dimensions, expressions juridiques, formules chimiques, nombres, indicateurs d'élements de listes). Pour cela, nous avons construit une architecture d'entraînement et d'étiquetage en plusieurs étapes.

À l'entraînement, le corpus est d'abord traité par notre chaîne d'analyse de surface SxPipe (cf. section 7.1.2), dans un mode qui préserve la segmentation d'origine en unités à annoter, de façon à reconnaître les mentions d'entités relevant des types cités ci-dessus. Les unités qui ne font pas partie d'une mention conservent leur annotation, alors que les mentions identifiées sont étiquetées par leur type. MElt est alors entraîné sur ce corpus modifié.

Lorsque l'on étiquette un nouveau corpus, la même configuration de SxPipe est appliquée et MElt est appliqué sur le résultat. Puisque l'on souhaite *in fine* étiqueter chacune des unités d'origine, il reste à étiqueter les unités qui constituent les mentions. L'étiquetage de l'intérieur des mentions relevant de conventions, nous avons développé pour cela une série d'heuristiques d'étiquetage spécifiques à chaque type d'entités et qui tentent de simuler au mieux les instructions du guide d'annotation [29]. On obtient donc une annotation en moyenne plus cohérente que le corpus d'évaluation lui-même, d'où des scores possiblement légèrement sous-estimés.

Malgré ce biais défavorable, les scores obtenus sont meilleurs que ceux présentés à la section précédente, comme le détaille le tableau 8.5 ($\text{MElt}_{\text{fr}}{}^{\text{SPMRL}}$+EN désigne l'architecture présentée dans cette section, alors que $\text{MElt}_{\text{fr}}{}^{\text{SPMRL}}$ est l'étiqueteur présenté à la section précédente).

Comme pour les expériences de la section précédente, nous nous sommes comparés à l'étiqueteur Marmot. Nous avons donc également encapsulé Marmot dans nos outils de gestion de certaines entités nommées pour comparer les gains obtenus par ce biais entre MElt et Marmot (colonnes « +EN »). Les résultats de ces expériences sont fournis au tableau 8.5, ainsi que les scores de TreeTagger sur FTB-SPMRL (avec et sans le L*efff* comme lexique externe), pour information.

Les résultats montrent une amélioration de MElt grâce à ce mécanisme qui est supérieure à celle obtenue avec Marmot. Ainsi, la précision de MElt, légèrement inférieure à celle de Marmot sans ce *wrapper*, devient très légèrement supérieure (de façon non

29. On pourrait procéder différemment, par exemple en construisant autant de corpus qu'il y a de type d'entités traitées, puis en entraînant un étiqueteur spécifique à chaque type de mention qui apprendrait à reproduire au mieux la façon dont l'intérieur de chaque type de mention est annoté dans les données d'entraînement. Cela éviterait le travail manuel de développement des heuristiques d'étiquetage mais souffrirait de l'incohérence des annotations à l'intérieur des mentions.

Système	Sans lexique externe				Avec lexique externe (Lefff)			
	standard		+EN		standard		+EN	
	Total	Inc.	Total	Inc.	Total	Inc.	Total	Inc.
TreeTagger	95,54	84,47			96,11	86,35		
Marmot	97,29	86,95	97,41	87,79	**97,41**	88,08	97,50	88,41
MElt	97,12	85,54	97,16	87,80	97,36	**88,16**	**97,51**	**90,63**

Tableau 8.5 – Évaluation comparative de l'exactitude de MElt, de Marmot et de TreeTagger (en %) sur le corpus FTB-SPMRL (33 étiquettes distinctes, 443 113 mots dans le corpus d'entraînement).

statistiquement significative). Sur les mots inconnus, les scores de MElt, déjà meilleurs que ceux de Marmot sans le *wrapper*, creusent l'écart lorsque ce dernier est utilisé (90,63% pour MElt contre 88,41% pour Marmot). Mais encore une fois, ces scores sont à prendre avec précaution étant donné le manque de cohérence dans la façon dont l'intérieur des mentions sont annotées dans les données d'évaluation.

8.2 Informations lexicales et étiquetage morphosyntaxique statistique : alVWtagger et la campagne d'évaluation UD 2017 [30]

Bien que MElt soit relativement rapide à entraîner, le développement d'étiqueteurs performants, notamment par l'exploration systématique de plusieurs hyperparamètres des modèles, finit par prendre un temps significatif lorsque l'on doit le faire en parallèle sur un grand nombre de jeux de données. Remplacer dans la procédure d'apprentissage le modèle à maximisation d'entropie par un perceptron [31] permet de diminuer les temps d'entraînement, mais au prix d'une légère baisse des performances [32].

Dans le cadre de notre participation (Villemonte de La Clergerie *et al.*, 2017) à la campagne d'évaluation multilingue des analyseurs syntaxiques organisée en marge de la conférence CoNLL 2017 dans le cadre de l'initiative *Universal Dependencies* (Zeman *et al.*, 2017) [33], nous avons dû développer des étiqueteurs morphosyntaxiques pour plusieurs dizaines de langues, et plus précisément pour un nombre encore plus grand de jeux de données. Nous nous sommes donc tournés vers l'architecture *open source* Vowpal Wabbit [34], conçue précisément pour optimiser les temps d'entraînement de modèles statistiques, et notamment linéaires comme utilisés par MElt. Nous avons ainsi développé

30. Ce travail est décrit dans la section correspondante de (Villemonte de La Clergerie *et al.*, 2017), article qui décrit notre participation à la campagne d'évaluation UD 2017 des analyseurs syntaxique, décrite ci-après.

31. Ce que permet MElt sur option, l'outil `megam` sur lequel il s'appuie permettant d'entraîner des modèles à maximum d'entropie comme des perceptrons mono- ou multiclasses.

32. Et au prix d'une certaine instabilité des résultats, dès lors, comme c'est le cas dans `megam`, que l'initalisation est aléatoire.

33. « Multilingual Parsing from Raw Text to Universal Dependencies », `http://universaldependencies.org/conll17/`.

34. `https://github.com/JohnLangford/vowpal_wabbit/wiki`

un nouvel étiqueteur morphosyntaxique, qui s'inspire de MElt tout en en améliorant encore le jeu de traits, et qui utilise Vowpal Wabbit comme moteur d'apprentissage.

8.2.1 La campagne d'évaluation UD 2017

Cette campagne d'évaluation, que nous dénoterons désormais par le terme « campagne UD 2017 » présentait quatre difficultés majeures :

— **Le caractère fortement multilingue et multidomaines**
Les jeux de données à analyser, dérivées de la version 2.0 de la collection *Universal Dependencies*, étaient au nombre de 81, couvrant 46 langues différentes, y compris quelques langues pour lesquelles aucun jeu de développement n'était fourni, ainsi que 4 langues « surprises » dont nous n'avons eu connaissance que quelques jours avant la période d'évaluation et pour lesquelles seuls des corpus annotés jouets de quelques centaines de mots ont été fournis.

— **L'analyse de corpus bruts**
Les campagnes d'évaluation SPMRL shared tasks 2013 et 2014 Seddah *et al.* (2013b, 2014) ont été les premières à mettre en jeu, pour deux des langues couvertes (l'hébreu et l'arabe), des configurations où la segmentation en unités lexicales et l'étiquetage morphosyntaxique étaient prédits, plutôt que de ne fournir que des données dont la segmentation et l'étiquetage étaient ceux des corpus de de référence. Dans le cas de la campagne CoNLL 2017, les organisateurs sont allés plus loin : les données à analyser étaient des données brutes, et les participants devaient donc segmenter en phrases, tokeniser, identifier les unités lexicales, procéder à l'étiquetage morphosyntaxique (parties du discours et, si nécessaire pour les analyseurs, traits morphologiques), et finalement à l'analyse syntaxique proprement dite. Les participants pouvaient toutefois faire usage des annotations fournies par le système de segmentation et étiquetage morphosyntaxique et morphologique UDPipe (Straka *et al.*, 2016) s'ils souhaitaient se concentrer uniquement sur l'analyse syntaxique.

— **L'inventaire fermé de ressources autorisées**
Pour développer leurs outils d'analyse, les participants n'avaient pas le droit d'utiliser n'importe quelle ressource. Ils devaient se limiter à une liste fermée qui incluait notamment des *word embeddings* (cf. section suivante) pré-entraînés et publiés par Facebook, les données parallèles des collections OPUS, les analyseurs morphologiques des projets Apertium et Giellatekno, et des données brutes issues de Wikipedia, ainsi que le résultat de leur analyse par la chaîne de traitement de surface UDPipe [35]. Il nous était donc impossible d'utiliser les lexiques Alexina, qu'il

35. Les autres ressources autorisées, dont nous ne nous sommes pas servis, étaient les données du *Word Atlas of Language Structures* [36] (WALS), les données parallèles de la campagne WMT 2016 d'évaluation des

s'agisse de ceux que nous avons développés ou des lexiques librement disponibles et que nous avons converti dans le formalisme Alexina (cf. chapitre 3).

— **L'environnement d'exécution des systèmes à l'aveugle**
Les systèmes devaient être installés sur la plateforme TIRA, sur lesquelles ils étaient exécutés à l'aveugle sur les données de test, sans qu'aucune information ne soit renvoyée aux participants avant la fin de la période d'évaluation, à l'exception d'une information indiquant que l'évaluation, très stricte quant au respect des caractères composant le texte source, s'était exécutée ou non avec succès [37].

8.2.2 alVWtagger

Nous nous concentrons ici sur le nouvel étiqueteur statistique alVWtagger que nous avons développé pour cette campagne d'évaluation. L'utilisation de Vowpal Wabbit nous a permis, grâce à la rapidité des entraînements, de tester pour chaque langue de nombreuses configurations différant par le lexique externe utilisé [38].

Outre un jeu de traits optimisé, alVWtagger se distingue de MElt de deux façons. La première concerne la façon dont les lexiques externes sont utilisées. Plutôt que d'utiliser seulement les catégories fournies par le lexique, nous utilisons aussi les traits morphologiques. Nous avons expérimenté à cet égard plusieurs configurations. Dans la configuration de base, les traits extraits du lexique font usage de la concaténation d'une partie du discours de l'inventaire des *Universal Parts-Of-Speech*, ci-après UPOS, et de la concaténation des traits morphologiques, eux aussi au format des *Universal*

systèmes de traduction automatique, et les jeux d'entraînement et de dévelopement d'une version du Corpus Arboré de Paris 7 convertie semi-automatiquement dans le modèle et le format UD 2.0.

37. Cette configuration nous a d'ailleurs valu une mésaventure, dont les rebondissements ont été relatés avec passion sur Twitter sous le mot-dièse #ParsingTragedy, mais qui n'a pas affecté nos résultats en étiquetage morphosyntaxique : la plateforme TIRA permettait de tester les systèmes sur les données de développement, mais également sur des données d'essai, qui étaient prévues pour être identiques quant à leur format aux données de test. Elles ne l'étaient pas. Nous utilisions, pour sélectionner le bon modèle d'analyse pour chaque jeu de données, une information dans un fichier de métadonnées qui était présente dans le jeu d'essai mais pas dans le jeu de test. Il en a résulté que notre sélectionneur de modèle ne trouvait pas les modèles entraînés pour chaque jeu de données, et qu'il a donc utilisé les modèles génériques, indépendants de la langue, selon un mécanisme que nous avions mis en place pour traiter les données des langues surprises. La configuration à l'aveugle de TIRA ne nous a pas permis de nous apercevoir du problème, nos résultats étant évalués avec succès. Même les organisateurs, qui informaient les participants lorsqu'ils obtenaient des scores trop bas pour être réalistes, n'ont pas détecté de problème, nos modèles génériques réussissant à produire des analyses d'apparence raisonnables, qui nous ont valu un classement de 27èmes sur 33 participants au classement principal, celui selon le score d'attachement étiqueté (*Labelled Attachment Score*, LAS). Ce n'est qu'après la fin de la période d'évaluation que nous avons pu avoir accès aux logs de nos analyseurs et nous apercevoir du problème. Nous avons corrigé notre sélectionneur de modèle et relancé notre système sans rien changer à nos modèles ou à notre chaîne de traitement, nous classant alors de façon semi-officielle 6èmes sur 33. Cette mésaventure et la semi-officialisation de ce classement font l'objet d'un paragraphe dédié dans l'article référence de Zeman *et al.* (2017) et est relatée plus en détails dans Villemonte de La Clergerie *et al.* (2017).

38. Dans toutes nos expériences, nous avons utilisé Vowpal Wabbit dans son mode multiclasse par défaut, c'est-à-dire avec une fonction de coût quadratique (et non pas logistique, ce qui aurait conduit à une architecture MEMM à l'image de MElt) et une stratégie *one-against-all*.

Features, ci-après UF. Ce mode est appelé « UPOS+UF ». Nous avons également testé des configurations où seule la UPOS est utilisées (mode « UPOS »), ainsi que des configurations où nous utilisons comme source de traits distincts soit le UPOS et la concaténation des UF (mode « UPOS/UF ») soit d'une part la catégorie totale UPOS+UF et d'autre part les UF (mode « UPOS+UF/UF »). Nous reviendrons ci-dessous sur la façon dont nous avons extrait et converti des lexiques morphologiques dans le schéma UPOS/UF uniquement à partir des ressources autorisées pour la campagne d'évaluation.

La deuxième amélioration par rapport à MElt est que alVWtagger peut prédire à la fois une catégorie (ici, une UPOS) et un ensemble de traits morphologiques (ici, des UF). Nous avons décidé de nous restreindre à un sous-ensemble de traits pour éviter les problèmes de dispersion des données produits par certains traits utilisés dans certains jeux de données [39] Pour chaque mot, alVWtagger prédit d'abord une UPOS, puis utilise cette UPOS parmi ses traits pour prédire de façon simultanée l'ensemble des UF, à l'aide d'un deuxième modèle. [40]

8.2.3 Extraction des lexiques morphologiques

Nous avons construit chaque fois que c'était possible plusieurs types de lexiques morphologiques à partir des ressources autorisées dans le cadre de la campagne UD 2017 [41]. Nous les avons systématiquement converties dans le format UPOS/UF au moyen de tableaux de correspondance développées manuellement. Nous avons comparé les performances de alVWtagger sur chaque jeu de données en faisant varier le type de lexique et la configuration (UPOS+UF, UPOS, UPOS/UF, UPOS+UF/UF) et en évaluant la performance en étiquetage UPOS sur le jeu de développement [42]. Les différents lexiques que nous avons utilisé sont issue de l'une ou de plusieurs des sources suivantes (après fusion *a posteriori* lorsque plusieurs sources sont utilisées) :

— Les lexiques monolingues du projet Apertium (code « AP » dans le tableau 8.6) ;

— Les données brutes fournies par les organisateurs de la campagne UD 2017, après application d'un tokeniseur de base, extraction du million de mots les plus fréquents, puis application des analyseurs morphologiques du projet Apertium ou, à défaut, du projet Giellatekno (codes « APma » or « GTma ») ;

— Le jeu d'entraînement des données de la campagne UD 2017 (code « T ») ou d'un autre jeu de données pour la même langue (code « T*jeu_de_données* ») ;

— Les données annotées par UDPipe fournies par les organisateurs (code « UDP ») ;

39. Nous avons retenu les traits suivants : Case, Gender, Number, PronType, VerbForm, Mood, and Voice.

40. Nous avons aussi testé des configurations où les traits morphologiques étaient prédits un par un, mais les résultats sur les jeux de développements étaient légèrement inférieurs.

41. En particulier, aucune des ressources utilisées à la section 8.1.3 n'était autorisée

42. Dans les cas où aucun jeu de développement n'était fourni, nous avons utilisé une technique de *jackknife* sur le jeu de test, découpé en 10 sous-parties.

— Un lexique préalablement extrait pour une autre langue, automatiquement « traduit » grâce à un algorithme dédié au moyen d'un lexique bilingue extrait des données OPUS, alignées au niveau des phrases (code « TR*source_language* »).

Pour certaines langues seulement, faute de temps, nous avons également créé des versions étendues de nos lexiques à l'aide d'embeddings que nous avons re-calculés à partir des données brutes fournies par les organisateurs au moyen de l'outil `word2vec`, en associant aux mots inconnus du lexique les informations associées au mot connu du lexique le plus proche selon une distance euclidienne dans l'espace des embeddings [43]. Lorsque le lexique conduisant aux meilleures performances est un lexique ainsi étendu, il est indiqué dans le tableau 8.6 par un suffixe « -e ».

Dans le cadre de la campagne UD 2017, nous avons choisi d'utiliser l'étiquetage produit par alVWtagger ou celui fourni par les organisateurs produit par UDPipe en fonction des performances en analyseur syntaxique qui en résultaient. Cela nous a permis d'arriver 3èmes sur 33 en étiquetage UPOS. Des expériences ultérieures, avec de meilleurs modèles d'analyse syntaxique, ont conduit à ce que l'étiquetage par alVWtagger soit utilisé sur un plus grand nombre de jeux de données, avec à la clé un classement virtuel comme 2èmes en étiquetage UPOS. Il est intéressant d'observer que c'est avec un système statistique et non neuronal que nous avons réussi à obtenir de telles performances. Cela pose la question, sur laquelle nous reviendrons à la question suivante, de la pertinence des approches neuronales pour cette tâche, approches qui donnent pourtant d'excellentes performances sur de nombreuses tâches, et parmi lesquelles des architectures come les LSTM (cf. plus bas) spécifiquement conçues avec en tête l'annotation de séquences. Cela confirme en tout cas la pertinence d'utiliser des ressources lexicales extraites, même dans un cadre contraint comme cela était le cas dans cette campagne d'évaluation, pour améliorer les performances en étiquetage morpholgique [44].

43. Nous n'avons pas utilisé les embeddings fournis car nous avons constaté empiriquement que la fenêtre de taille 10 utilisée pour calculer ces embeddings conduisait à des résultats moins bons qu'en utilisant une fenêtre plus petite, en particulier lorsque les corpus bruts étaient de taille modeste.

44. À titre d'illustration, le lexique que nous avons pu extraire pour le Kazakh à partir d'un corpus brut et de l'analyseur morphologique d'Apertium nous a permis d'être classé premier en étiquetage UPOS pour cette langue, avec une précision de 67,86%, contre 58,48% pour le second, l'équipe de Standford par ailleurs grand vainqueur de cette campagne d'évaluation. Nous avons également retiré une satisfaction particulière de nos scores sur le corpus de slovaque, langue pour laquelle nous n'avons pas eu d'autre choix que de créer un lexique morphologique par traduction du lexique tchèque extrait du projet Apertium, via la stratégie esquissée plus haut. Ce lexique nous a en effet permis d'arriver deuxièmes en étiquetage UPOS, avec un score de 95,10% contre 96,87% pour le premier (Stanford) et devant l'équipe de l'IMS (Stuttgart), troisièmes avec un score de 94,60%.

Jeu de données	Langue	alVWtagger			UDPipe
		type de lexique	ms mode	précision globale	précision globale
ar	Arabe	AP-e	M	**94,71**	94,57
bg	Bulgare	AP	F	97,61	**97,72**
ca	Catalan	AP-e	FM	**98,42**	98,15
cs	Tchèque	AP	M	**98,83**	98,48
cs_{cac}	Tchèque	Tcs		**99,24**	98,78
cs_{cltt}	Tchèque	AP+Tcs	F	**94,34**	92,06
cu	Vieux-slave	T	F	**95,15**	94,07
da	Danois	AP		**96,30**	95,19
de	Allemand	AP	M	**92,70**	91,39
el	Grec moderne	AP	F	**95,53**	94,17
en	$Anglais_{web}$	AP	F	**94,68**	94,43
en_{lines}	Anglais	AP	FM	**96,08**	94,75
en_{partut}	Anglais	AP+T	FM	**95,90**	94,39
es	$Espagnol_{web}$	AP		**96,47**	96,24
es_{ancora}	Espagnol	AP	FM	**98,39**	98,16
et	Estonien	GTms	FM	**89,28**	87,52
eu	Basque	AP	F	**94,48**	92,80
fa	Persan	*pas de lexique*		96,04	**96,17**
fi	Finnois	GTms	FM	**95,06**	94,52
fi_{ftb}	Finnois	GTms	F	**92,50**	92,34
fr	$Français_{web}$	AP-e		**97,30**	97,08
$fr_{sequoia}$	Français	AP-e	FM	**97,54**	96,60
ga	Irlandais	UDP	M	—	—
gl	Galicien	AP	FM	**97,45**	96,77
got	Gothique	T		**94,53**	94,22
grc	Grec ancien	$Tgrc_{proiel}$-e		**89,56**	81,54
grc_{proiel}	Grec ancien	UDP-e	FM	**96,40**	96,01
he	Hébreu	AP	FM	**96,68**	95,72
hi	Hindi	AP	F	**96,59**	95,79
hr	Croate	TRsl	M	**96,94**	96,15
hu	Hongrois	T		**93,90**	92,31
id	$Indonésien_{web}$	AP	M	92,98	**93,36**
it	Italien	AP	F	**97,55**	97,23
it_{partut}	Italien	Trit	M	**97,89**	95,16
ja	Japonais	*pas de lexique*		**96,87**	96,72
kk	Kazakh	APms		—	—
ko	Coréen	*pas de lexique*		**93,77**	93,68
la_{ittb}	Latin	TRit+T		**97,15**	96,86
la_{proiel}	Latin	TRit+T-e	FM	**95,62**	95,43
lv	Letton	AP	FM	**93,43**	90,81
nl	Néerlandais	AP	F	**94,70**	94,07
$nl_{lassysmall}$	Néerlandais	AP+Tnl	F	**96,74**	95,65
$no_{bokmaal}$	Norvégien (Bokmål)	AP		**97,66**	97,34
$no_{nynorsk}$	Norvégien (Nynorsk)	AP	M	**97,23**	96,74
pl	Polonais	AP	M	**97,03**	95,34
pt	Portugais	AP	FM	**97,21**	97,00
pt_{br}	Portugais (Brésil)	AP+Tpt	FM	**97,96**	97,40
ro	Roumain	AP		**97,34**	96,98
ru	Russe	AP	M	**96,62**	94,95
$ru_{syntagrus}$	Russe	AP	FM	**98,54**	98,20
sk	Slovaque	TRcs	FM	**96,00**	93,14
sl	Slovène	AP	FM	**97,82**	96,34
sv	Suédois	AP	FM	**96,32**	95,17
sv_{lines}	Suédois	AP	F	**96,01**	94,63
tr	Turc	APma	FM	**93,65**	92,25
ug	Ouïghour	UDP		—	—
uk	Ukrainien	AP	M	—	—
ur	Ourdou	AP		**93,01**	92,45
vi	Vietnamien	*pas de lexique*		88,60	**88,68**
zh	Chinois (Mandarin)	AP+UDP		**91,40**	91,21

Tableau 8.6 – Précision des meilleurs modèles alVWtagger en étiquetage UPOS sur les données de développement de la campagne d'évaluation UD 2017, comparée aux résultats du système UDPipe.

8.3 Informations lexicales et étiquetage morphosyntaxique neuronal : alNNtagger [45]

Depuis quelques années, et comme nous venons de l'évoquer, les approches reposant sur des réseaux de neurones font l'objet d'un renouveau important et ont permis d'améliorer l'état de l'art, parfois de façon très significative, pour un grand nombre de tâches de traitement automatique des langues. Dans ce contexte, nous avons cherché à répondre à deux questions. La première est de savoir si cela s'applique à la tâche d'étiquetage morphosyntaxique, comme le font notamment penser les travaux de Plank *et al.* (2016). La seconde est de comprendre comment intégrer les informations lexicales issues d'un lexique externe à un étiqueteur neuronal. Dans cette section, nous nous concentrons avant tout sur la seconde de ces deux questions, tout en discutant brièvement la première en conclusion.

Dans une approche neuronale, le rôle et l'utilité des informations lexicales issues d'un lexique externe est une question d'autant plus intéressante qu'une autre source d'informations lexicales peut être utilisée de façon directe, à savoir les représentations vectorielles des « mots » appelées *embeddings* (Bengio *et al.*, 2003 ; Collobert et Weston, 2008 ; Chrupała, 2013 ; Ling *et al.*, 2015 ; Ballesteros *et al.*, 2015 ; Müller et Schütze, 2015). De telles représentations, que nous appellerons désormais *w-embeddings*, sont utilisées comme représentation d'entrée dans les réseaux de neurones, qui les optimisent au cours de l'entraînement. Il est toutefois possible d'initialiser cette couche d'entrée au moyen de w-embeddings extraits à partir de volumes importants de données textuelles brutes, ce qui constitue une autre source d'informations lexicales externes. De tels w-embeddings pré-entraînés se sont révélés utiles pour de nombreuses tâches. C'est notamment le cas pour l'étiquetage morphosyntaxique, notamment dans des architectures reponsant sur les réseaux neuronaux récurrents (RNN) et, plus spécifiquement, sur les réseaux à longue mémoire à court terme (LSTM) de niveau mot ou de niveau caractère, mono- ou bi-directionnels (Hochreiter et Schmidhuber, 1997 ; Ling *et al.*, 2015 ; Ballesteros *et al.*, 2015 ; Plank *et al.*, 2016).

Les *embeddings* construits au niveau des caractères, ci-après *c-embeddings* [46] sont particulièrement pertinents pour l'étiquetage morphosyntaxique en ceci qu'ils constituent des représentations vectorielles qui encodent la séquence de caractères constitutive de chaque « mot ». Ils peuvent ainsi encoder des généralisations pertinentes sur des sous-parties telles que des préfixes ou des suffixes, ce qui est utile pour traiter les mots inconnus du corpus d'apprentissage. Toutefois, tous ne sont pas morphologiquement réguliers, et

45. Le travail présenté dans cette section, réalisé en collaboration avec Héctor Martínez Alonso, post-doctorant à ALMAnaCH sous ma supervision, a fait l'objet d'une publication (Sagot et Martínez Alonso, 2017).

46. À ne pas confondre avec les embeddings de caractères, qui construisent des représentations pour chaque caractère.

dans de tels cas, les c-embeddings ne peuvent rien apporter, et pourraient même avoir une influence négative sur les performances. C'est là une différence importante avec les lexiques externes, qui fournissent des informations pertinentes sur tous les mots qu'ils contiennent. À l'inverse, les lexiques externes ne distinguent pas quantitativement les informations pertinentes de celles qui le sont moins.

Nous avons donc mené une évaluation comparative des avantages respectifs de l'utilisation de c-embeddings et de lexiques externes et l'impact de leur utilisation conjointe dans un étiqueteur morphologique neuronal. Nous nous sommes appuyés pour cela sur l'étiqueteur de Plank *et al.* (2016), pour lequel nous montrons que l'intégration de l'information issue de lexiques externes permet une amélioration des performances, y compris lorsque l'on utilise des c-embeddings, et ce même lorsque les w-embeddings sont initialisés au moyen de w-embeddings appris sur des corpus bruts.

8.3.1 Étiquetage par bi-LSTM et intégration de l'information lexicale

Comme l'a montré Plank *et al.* (2016) avec des expériences sur plusieurs dizaines de jeux de données représentant autant de langues différentes, des performances de niveau état de l'art peuvent être atteintes au moyen d'une architecture bi-LSTM (LSTM bi-directionnels) de niveau mot. Les meilleures performances sont obtenues en utilisant à la fois des w-embeddings et des c-embeddings, c'est-à-dire en représentant chaque mot par la concaténation d'un w-embedding $\vec{w}$ construit par une couche d'entrée dédiée et d'un c-embedding construit au moyen d'un bi-LSTM de niveau caractère entraîné conjointement. Une amélioration supplémentaire des performances est possible sur certains jeux de données, mais pas sur tous, en initialisant la couche de w-embedding par des w-embeddings pré-calculés sur des corpus externes (cf. Plank *et al.*, 2016 pour plus de détails).

Nous avons donc cherché à étendre cette architecture bi-LSTM en complétant cette représentation des mots par un vecteur $\vec{l}$ extrait d'un lexique externe. Pour un mot donné, ce vecteur $\vec{l}$ est une représentation dite « *n*-hot », c'est-à-dire un vecteur booléen dont chaque valeur correspond à l'une des étiquettes possibles fournies par le lexique : la valeur est mise à 1 si et seulement si le lexique associe la catégorie correspondante au mot courant. Par exemple, le mot anglais *house*, qui peut être un nom au singulier ou un verb dans sa forme de base, sera associé à un vecteur dont seules deux coordonnées sont à 1, correspondant aux deux catégories possibles selon le lexique. Les mots inconnus du lexique sont représentés par un vecteur dont toutes les coordonnées sont à 0. On notera que, tout comme lorsque le lexique externe est utilisé sous forme de traits dans un modèles tel que celui sur lequel repose MElt, il n'est pas nécessaire que l'inventaire des catégories du lexique soit le même que celui du corpus d'entraînement.

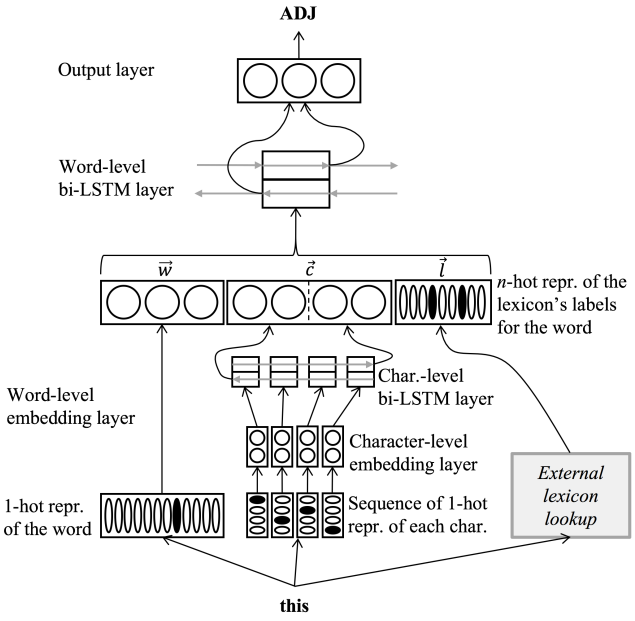


FIGURE 8.2 – Schéma de notre extension de l'architecture bi-LSTM d'étiquetage de Plank *et al.*'s (2016) afin d'y intégrer l'expoitation d'informations provenant d'un lexique externe. Ce schéma représente le traitement d'un mot anglais unique, « this ». Les connections des cellules LSTM de niveau mot à ses contreparties pour les mots précédent et suivant sont représentées par des flèches grises.

La figure 8.2 illustre la façon dont ce vecteur $\vec{l}$ est intégré par concaténation aux représentations sous forme de w-embedding ($\vec{w}$) et de c-embedding ($\vec{c}$). C'est le résultat de ce cette concaténation qui est donné en entrée à la couche bi-LSTM.

8.3.2 Données

Pour faciliter la comparaison avec (Plank *et al.*, 2016), nous avons mené nos expériences sur les mêmes corpus, à savoir les corpus de la version 1.3 des *Universal Dependencies* (Nivre *et al.*, 2016).

8.3.2.0.1 Lexiques externes Nous avons mené des expériences préliminaires avec un à trois lexiques pour chaque langue, constuits pour certains à partir de ressources Alexina (natives ou converties) comme évoqué à la section 8.1.3 et pour d'autres à partir des données distribuées par les projets Apertium et Giellatekno, comme décrit à la section 8.2.3 (quoique dans une version légèrement antérieure). Dans le cas des lexiques extraits d'Apertium et de Giellatekno, nous en avons systématiquement produit deux variantes. La première, dite « à gros grain », utilise pour chaque entrée son *Universal Category* (UPOS) comme catégorie. La seconde, « à grain fin », utilise pour chaque entrée la concaténation de son UPOS et de ses *Universal Features* (UF). Ces deux versions du

même lexique correspondent exactement aux configurations UPOS et UPOS+UF de la section 8.2.3.

Pour chaque jeu de données, nous avons choisi l'un des un à trois lexiques ainsi construits en ne retenant que celui consuisant à des performances maximales sur le jeu de développement. C'est ce que nous appelons le « meilleur » lexique. Dans le reste de cette section, tous les résultats ont été obtenus en utilisant pour chaque jeu de données le « meilleur » lexique correspondant. Le tableau 8.7 et sa légende fournissent des informations complémentaires sur ces « meilleurs » lexiques.

8.3.2.0.2 W-embeddings pré-calculés À l'image des expériences de Plank *et al.* (2016), nous avons également utilisé les w-embeddings pré-calculés Polyglot Al-Rfou *et al.* (2013). Les langues (et donc les jeux de données) pour lesquelles des w-embeddings Polyglot sont disponibles sont indiquées dans le tableau 8.7.

Nous avons entraîné notre étiqueteur neuronal à la fois avec et sans c-embeddings, avec et sans initialisation grâce aux w-embeddings Polyglot (pour les cas où ils étaient disponibles), et avec et sans utilisation de notre mécanisme d'intégration d'informations lexicales externes. Il en résulte donc entre 4 et 12 configurations pour chaque jeu de données.

8.3.3 Expériences

Le système dont nous sommes partis, et qui nous sert de base de comparaison, est `bilty` [46]. Il s'agit d'une version « refactorisée » et librement disponible du code d'origine utilisé par Plank *et al.* (2016). Nous utilisons sa configuration standard, avec une couche bi-LSTM unique, des c-embeddings de dimension 100, des w-embeddings de dimension 64 (la même que les w-embeddings de Polyglot), pas de tache annexe [47] et 20 iterations pour l'entraînement. Nous avons étendu `bilty` pour qu'il puisse également prendre en compte des informations lexicales externes de la façon décrite ci-dessus.

Pour chaque configuration (avec ou sans lexique externe), nous avons entraîné trois versions de l'étiqueteur : (i) une version sans c-embeddings et sans initialisation par les w-embeddings Polyglot, (ii) une version avec c-embeddings et sans initialisation par les w-embeddings Polyglot, et (iii) lorsque des w-embeddings étaient disponibles, une version avec c-embeddings et initialisation de la couche de w-embedding par les w-embeddings

46. https ://github.com/bplank/bilstm-aux

47. La tâche annexe utilisée par Plank *et al.* (2016), à savoir prédire la classe de fréquence de chaque mot, produit de meilleurs scores sur les mots inconnus mais des scores globaux identiques en macro-moyenne à la configuration sans apprentissage multi-tâche.

Jeu de données	Langue	Lexique	#entrées ($\times 10^3$)	#étiqu.	couv. (%)	type-token ratio	Polyglot
ar	Arabe	AP	651	15	46	0,09	oui
bg	Bulgare	Multext-East	53	12	68	0,18	oui
ca	Catalan	AP	379	13	80	0,06	oui
cs	Tchèque	AP	1 875	15	68	0,10	oui
da	Danois	AP	683	15	72	0,19	oui
de	Allemand$_{\text{web}}$	DeLex	465	52	79	0,18	oui
el	Grec	AP	47	12	48	0,20	oui
en	Anglais$_{\text{web}}$	AP	127	12	78	0,09	oui
es	Espagne	Le*ffe*	756	34	87	0,12	oui
et	Estonien	GTma	44	12	52	0,23	oui
eu	Basque	AP$_{\text{full}}$	53	14	42	0,22	oui
fa	Persan	PerLex	512	37	69	0,10	oui
fi	Finnois	GTma	228	13	54	0,29	oui
fr	Français$_{\text{web}}$	Le*fff*	539	25	88	0,11	oui
ga	Irlandais	inmdb	114	32	35	0,26	oui
gl	Galicien	AP	241	12	85	0,12	non
grc	Grec ancien	Diogenes	1,314	18	46	0,20	non
he	Hébreu	AP	268	16	71	0,12	oui
hi	Hindi	AP	159	14	83	0,05	oui
hr	Croate	HML	1 361	22	85	0,21	oui
id	Indonésien$_{\text{web}}$	AP$_{\text{full}}$	12	38	61	0,18	non
it	Italien	AP	278	14	78	0,10	oui
kk	Kazakh	APma	434	16	79	0,48	non
la	Latin	Diogenes	562	16	83	0,31	non
lv	Letton	AP	314	14	58	0,33	non
nl	Néerlandais	Alpino	81	65	77	0,14	oui
no	Norvégien	AP	2 470	13	66	0,11	oui
pl	Polonais	AP	1 316	15	61	0,31	oui
pt	Portugais	AP	159	155	72	0,13	oui
ro	Roumain	Multext-East	378	14	66	0,18	non
ru	Russe	AP	4 401	16	61	0,32	non
sl	Slovène	AP	654	14	70	0,24	oui
sv	Suédois	Saldo	1 215	214	88	0,17	oui
tr	Turc	APma	417	14	74	0,32	no
zh	Chinois	AP	8	13	34	0,16	non

Tableau 8.7 – Informations sur les données lexicales utilisées dans nos expériences sur alNNtagger. Le « meilleur » lexique est indiqué avec sa taille, la taille de son jeu d'étiquettes et sa couverture sur le sous-corpus de test. Les lexiques extraits d'Apertium ou de Giellatekno sont spécifiés par les mêmes codes que dans le tableau 8.6. Il s'agit toujours de leur variante « à gros grain », sauf lorsque l'indice *full* est présent. Sont également fournis le rapport entre le nombre de tokens distincts et le nombre de tokens total (TTR) dans le corpus d'entraînement et l'indication de la disponibilité de w-embeddings pré-entraînés Polyglot (PG) utilisés pour initialiser la couche de w-embedding ($\vec{w}$).

Jeu de données	Système de base (sans lexique)			Avec le « meilleur » lexique externe (cf. tab. 8.7)			Écart avec le système de base		
	$\vec{w}$	$\vec{w}+\vec{c}$	$\vec{w}_P+\vec{c}$	$\vec{w}+\vec{l}$	$\vec{w}+\vec{c}+\vec{l}$	$\vec{w}_P+\vec{c}+\vec{l}$	$\vec{w}(+\vec{l})$	$\vec{w}+\vec{c}(+\vec{l})$	$\vec{w}_P+\vec{c}(+\vec{l})$
ar	93,90	95,99	96,20	94,58	96,05	96,22	+0,68	+0,06	+0,02
bg	94,50	98,11	97,62	96,29	98,30	97,86	+1,79	+0,18	+0,24
ca	96,14	98,03	98,17	97,58	98,21	98,26	+1,44	+0,18	+0,09
cs	95,93	98,03	98,10	96,74	98,46	98,41	+0,81	+0,43	+0,31
da	90,16	95,41	95,62	94,20	96,24	96,14	+4,04	+0,83	+0,53
de	87,94	92,64	92,96	91,52	93,08	93,18	+3,58	+0,44	+0,23
el	95,62	97,76	98,22	96,03	97,67	98,17	+0,41	−0,09	−0,05
en	91,12	94,38	94,56	92,97	94,63	94,70	+1,85	+0,25	+0,14
es	93,10	94,96	95,27	94,62	94,84	95,07	+1,52	−0,11	−0,20
et	90,73	96,10	96,40	90,07	96,14	96,66	−0,65	+0,04	+0,26
eu	88,54	94,34	95,07	88,52	94,78	95,03	−0,02	+0,44	−0,04
fa	95,57	96,39	97,35	96,22	97,09	97,35	+0,65	+0,71	+0,00
fi	87,26	94,84	95,12	88,67	94,87	95,13	+1,40	+0,03	+0,01
fr	94,30	95,97	96,32	95,92	96,71	96,28	+1,62	+0,74	−0,04
ga	86,94	89,87	91,91	88,88	91,18	91,76	+1,94	+1,31	−0,16
gl	94,78	96,94	—	95,72	97,18	—	+0,94	+0,24	—
grc	88,69	94,40	—	89,76	93,75	—	+1,07	-0,65	—
he	92,82	95,05	96,57	94,11	95,53	96,76	+1,29	+0,48	+0,19
hi	95,55	96,22	95,93	96,22	96,50	96,95	+0,67	+0,28	+1,02
hr	86,62	95,01	95,93	93,53	96,29	96,34	+6,91	+1,28	+0,41
id	89,07	92,78	93,27	91,17	92,79	92,89	+2,11	+0,02	−0,38
it	95,29	97,48	97,77	97,54	97,81	97,88	+2,26	+0,33	+0,11
kk	72,74	76,32	—	82,28	82,79	—	+9,54	+6,47	—
la	85,18	92,18	—	90,63	93,29	—	+5,44	+1,12	—
lv	78,22	89,39	—	83,56	91,07	—	+5,35	+1,68	—
nl	84,91	89,97	87,80	85,20	90,69	89,85	+0,29	+0,72	+2,05
no	93,65	97,50	97,90	95,80	97,72	97,96	+2,15	+0,22	+0,07
pl	87,99	96,21	96,90	90,81	96,40	97,02	+2,83	+0,18	+0,13
pt	93,61	97,00	97,27	94,76	96,79	97,11	+1,15	−0,21	−0,16
ro	92,63	95,76	—	94,49	96,26	—	+1,86	+0,51	—
ru	84,72	95,73	—	93,50	96,32	—	+8,79	+0,60	—
sl	83,96	97,30	95,27	94,07	97,74	95,44	10,11	+0,44	+0,17
sv	92,06	96,26	96,56	95,61	97,03	97,00	+3,55	+0,77	+0,44
tr	87,02	93,98	—	90,03	93,90	—	+3,01	−0,08	—
zh	89,17	92,99	—	89,29	93,04	—	+0,12	+0,05	—
Macro-moyenne	90,01	94,61	—	92,60	95,18	—	+2,59	+0,57	—
id. avec embed. PG	91,43	95,52	95,77	93,52	95,91	95,98	+2,09	+0,38	+0,21

Tableau 8.8 – Résultats globaux d'alNNtagger sur les corpus UD 1.3. Les configurations de base (sans lexique externe) correspondent à `bilty` (Plank *et al.*, 2016). Les scores de précision sont données pour les configuration : avec w-embeddings uniquement ($\vec{w}$), avec c-embeddings et w-embeddings ($\vec{w}+\vec{c}$), et avec c-embeddings et w-embeddings initialisés grâce aux w-embeddings pré-calculés fournis par Polyglot (PG) ($\vec{w}_P+\vec{c}$). Les dernières colonnes montrent l'écart entre les configurations avec lexique externe et les configurations de base.

Polyglot. Ce protocole est délibérément similaire à celui de Plank *et al.* (2016), afin de permettre une comparaison directe des résultats. [48]

8.3.4 Résultats et discussion

Les résultats montrent que l'utilisation d'informations issues de lexiques externe résulte en une amélioration quasi-systématique des résultats sur notre ensemble de 35 jeux de données couvrant autant de langues. Sans surprise, les améliorations les plus significatives sont obtenues lorsque l'on n'utilise pas de c-embeddings, avec un écart macro-moyenné de +2.56, contre +0.57 lorsque l'on utilise aussi les c-embeddings. L'utilisation des w-embeddings Polyglot pour initialiser la couche de w-embedding conduit à une amélioration légèrement plus faible, mais bien réelle.

Les améliorations observées sont particulièrement significatives pour les jeux de données dont les corpus d'entraînement sont plus petits : dans le configuration $\vec{w} + \vec{c}$, les trois jeux de données pour lesquels l'adjonction de $\vec{l}$ conduit aux améliorations les plus importantes sont également ceux dont les données d'entraînement sont les moins volumineuses.

On peut ainsi penser que les informations fournies par les lexiques externes restent pertinents même dans ce type de configuration en tant qu'elles couvrent explicitement les mots dont la morphologie est irrégulière et un certain nombre de mots inconnus du corpus d'entraînement, voire du corpus brut utilisé par Polyglot pour produire les w-embeddings pré-calculés. Autrement dit, les différents types d'embeddings ne répondent pas complètement au problème posé par les mots inconnus. Mais une étude approfondie de l'interaction entre les différentes sources d'informations que sont les w-embeddings, les c-embeddings et le lexique externe reste à faire. De même, des architectures plus complexes pourraient être testées, par exemple en utilisant plusieurs couches bi-LSTM ou une couche d'embedding sur le vecteur *n*-hot produit par le lexique externe.

Ce travail devra également être complété par d'autres expériences faisant usage de jeux d'étiquettes plus à granularité plus fine et donc de taille plus importante. À cet égard, Horsmann et Zesch (2017) ont montré récemment, en s'appuyant là aussi sur l'étiqueteur de Plank *et al.* (2016), que, sur des jeux d'étiquettes plus riches, cette architecture LSTM (sans utilisation de lexique externe, dans leur cas) se comporte moins bien que d'autres architectures non-neuronales mais faisant usage de lexiques externes [49]. Ceci dit, il reste à

48. Nous n'avons pas pris en compte les corpus alternatifs de la collection UD 1.3 (par exemple `nl_lassysmall` vs. `nl`), pas plus que les corpus pour des langues pour lesquelles nous n'avions ni w-embeddings Polyglot ni lexique externe (vieux slave, hongrois, gothique, tamoul).

49. Horsmann et Zesch (2017) écrivent ainsi : « Cependant, nous observons également que pour de très grands jeux d'étiquettes pour des langues à morphologie riche, des lexiques morphologiques développés manuellement [sic] sont toujours nécessaires pour atteindre des performances de niveau état-de-l'art » (*However, we also find that for the very large tagsets of morphologically rich languages, hand-crafted morphological lexicons are still necessary to reach state-of-the-art performance*).

comparer les performances de notre architecture neuronale avec lexique externe à celles d'étiqueteurs plus classiques, tels que MElt, faisant également usage d'un lexique externe. Après tout, si les approches neuronales ont le vent en poupe, et ce pour de bonnes raisons, leur supériorité n'est peut-être pas systématique, surtout sur une tâche comme l'étiquetage morpho-syntaxique pour laquelle les traits pertinents sont relativement simples à identifier manuellement. À cet égard, les résultats que nous avons obtenus lors de la campagne d'évaluation UD 2017 semblent indiquer qu'il est possible d'être très performant avec des modèles statistiques non-neuronaux, notamment lorsque les données d'apprentissage sont de volume modeste. Des expériences préliminaires que nous avons menées mais que nous ne rapporterons pas ici semblent confirment cette intuition, puisque alNNtagger et alVWtagger y ont montré des performances comparables en moyenne en étiquetage UPOS sur une dizaine de jeux de données de la collection de corpus UD 1.3.

8.4 Étiquetage morphosyntaxique de corpus bruts [50]

Comme nous l'avons fait depuis le début de ce chapitre, évaluer un étiqueteur morphosyntaxique consiste simplement à mesurer son taux d'exactitude, c'est-à-dire le pourcentage de « mots » ayant reçu la bonne étiquette. Ce pourcentage est souvent complété par l'exactitude sur les seuls mots inconnus (absents du corpus d'apprentissage). Toutefois, il faut garder à l'esprit que les comparaisons que l'ont peut faire de cette façon entre différents systèmes n'évaluent pas vraiment les systèmes en soi — quoi que cela puisse signifier —, mais plutôt leur adéquation au corpus utilisé, avec toutes ses caractéristiques et notamment sa taille, son homogénéité ou encore les caractéristiques de son jeu d'étiquettes (nombre d'étiquettes distinctes…).

Par ailleurs, comme nous venons de le voir, un étiqueteur morphosyntaxique est destiné à être utilisé *in fine* sur des corpus bruts, ce qui nécessite un découpage préalable en unités élémentaires destinées à recevoir des étiquettes, comme nous avons tenté de le faire lors de la campagne UD 2017. Dans les expériences présentées aux sections 8.1 à 8.3, comme souvent dans les travaux en étiquetage morphosyntaxique, cette problématique est éludée par l'utilisation de corpus de test déjà segmentés en « phrases » et en unités élémentaires à étiqueter. On peut considérer que de telles unités élémentaires, en tant qu'elles seront associées à des étiquettes linguistiquement signifiantes, doivent être des unités linguistiquement valides et donc correspondre à des unités lexicales.

Mais la détection de telles unités à partir de textes bruts une tâche difficile, comme nous l'avons vu au chapitre précédent. Outre la segmentation en « phrases », rappelons

50. Le travail présenté dans cette section, bien que de notre seul fait, a été réalisé dans cadre de travaux réalisés en collaboration, notamment avec Djamé Seddah. Il fait l'objet de certaines parties de plusieurs publications (Seddah *et al.*, 2012c,d,b ; Chanier *et al.*, 2014).

les trois difficultés majeures identifiées au chapitre précédent : (i) l'identification des composés et des amalgames, tâche particulièrement difficile pour des langues comme le chinois ou le thai (pas de séparateur typographique) ou encore le sanskrit (impact massif de phénomènes de *sandhi*), mais reste délicate pour l'ensemble des langues ; (ii) l'identification des entités nommées, qu'il s'agisse des entités nommées au sens strict (noms de lieux, de personnes, d'organisations) ou au sens plus large (dates, adresses, URL, mesures, etc.) ; (iii) l'orthographe et la typographie non-standard que l'on trouve dans certains types de corpus, dont les corpus issus du web.

Sauf à traiter l'ensemble de ces problèmes de façon jointe, comme nous l'avons réalisé de façon rudimentaire dans le cadre de la campagne UD 2017 [51], il est nécessaire, et *de facto* fréquent, de traiter ces difficultés par des approximations dont l'impact peut être important et dont la pertinence dépend de la langue considérée. Citons-en quelques-unes, qui ne sont nullement incompatibles entre elles.

Approximation 1 La première des approximations possibles consiste à ramener la tâche de découpage en unités élémentaires à une tâche de découpage en tokens : bien que les tokens soient des unités non linguistiques, on considère alors qu'ils sont suffisamment proches d'unités linguistiques propres à recevoir des annotations morphosyntaxiques pour que l'on puisse considérer qu'il est raisonnable d'annoter morphosyntaxiquement des tokens. Les conséquences de cette approximation sont par ailleurs limitées en partie par le fait que les composés ont souvent des composants identifiables comme relevant d'une catégorie connue [52]. Naturellement,

51. Le travail que nous avons fait à cet égard nous a permis d'atteindre de bons résultats sur un nombre non négligeable de jeux de données de la campagne UD 2017 sur les tâches de segmentation en phrases et de tokenisation. Précisons que la notion de token dans les données UD ne correspond pas exactement, pour les systèmes d'écriture sans séparateur typographique pertinent (chinois, japonais, vietnamien), à notre propre définition : pour simplifier, les tokens UD correspondent dans ces cas-là à des formes. L'idée que nous avons mise en œuvre est la suivante. Nous avons tout d'abord utilisé un outil de pré-tokenisation, dont l'objectif est d'avoir un rappel aussi élevé que possible pour la détection de frontières, qui sont à ce stade de nature sous-spécifiées (frontières de tokens ou frontières de phrases). Par exemple, en l'absence de séparateur typographique, chaque caractère est un pré-token. Nous avons alors utilisé un dérivé d'alVWtagger, que nous avons nommé alVWtokeniser, pour associer à chaque frontière de pré-tokens une étiquette parmi le jeu suivant : « frontière de token », « frontière de phrase », « pas une frontière ». Nous avons préservé pour chaque pré-token l'information selon laquelle cette dernière se situe ou non au niveau d'un caractère d'espace. Ainsi, par exemple, une étiquette « pas une frontière » au niveau d'un espace correspond à un « *token with spaces*. Nous avons entraîné alVWtokeniser sur les données d'entraînement fournies en lui fournissant les mêmes lexiques externes, et nous l'avons utilisé dès lors qu'il avait de meilleurs résultats sur les données de développement que la segmentation fournie par la *baseline* UDPipe (dans certains cas, notre segmentation en phrases était moins bonne mais notre segmentation en tokens était meilleure ; nous avons alors conservé la segmentation en phrases d'UDPipe et demandé à alVWtokeniser de ne procéder qu'à la segmentation en tokens). Cette stratégie, sur laquelle nous revenons plus en détails dans (Villemonte de La Clergerie *et al.*, 2017), nous a permis d'être classés 5èmes sur la tâche de tokenisation (précision de 98,85% ; les écarts sont faibles avec les autres participants : les deux premiers sont à 98,95%, la baseline UDPipe à 98,77) et 6èmes sur la tâche de segmentation en phrase (88,61% ; ici aussi les écarts sont faibles : le 4ème est à 88,68%, le 1er à 89,10, la baseline UDPipe à 88,49).

52. Ainsi, *pomme de terre* n'est pas difficile à analyser comme une séquence nom+préposition+nom, bien qu'il s'agisse en réalité d'un nom composé.

une telle approche nécessite que les corpus annotés sur lesquels les modèles sont entraînés et évalués reflètent ce même choix. C'est le cas par exemple dans le cadre des campanges d'évaluation SMPRL 2013 et 2014 ainsi que dans la campagne d'évaluation UD 2017 que nous avons évoquée à la section 8.2.

Approximation 2 Une approximation moins forte que la première consiste à ne pas ignorer les entités nommées, et donc à les reconnaître comme telles et à les étiqueter comme des unités, mais à se ramener au cas précédent pour tous les autres tokens. Il est pourtant rare que les données d'évaluation soient compatibles avec un tel modèle d'annotation, et on est alors obligé de disposer de stratégies complémentaires pour étiqueter les tokens qui font partie des mentions d'entités nommées. *In fine*, on est donc ramené à l'approximation précédente, bien que certaines étapes du traitement ait permis de l'éviter temporairement. C'est ce que nous avons tenté à la section 8.1.4.

Approximation 3 Une autre approximation consiste à ramener la tâche d'identification des formes à une tâche de correction orthographique des formes mal orthographiées ou mal tokenisées, notamment dans des corpus issus du web. La correction est alors effectuée de telle façon que les unités produites en sortie du correcteur ressemblent le plus possible à des formes, quand bien même elles sont identifiées ensuite comme s'il s'agissait de tokens. C'est l'objet de la présente section, qui s'appuie sur la méthodologie de correction automatique décrite à la section 7.3.3.

8.4.1 Méthodologie pour l'annotation morphosyntaxique de textes bruités par normalisation temporaire

Nous avons évoqué à la section 7.3.3 l'une des méthodologies de correction lexicale que nous avons utilisées, méthodologie qui repose simplement sur des règles développées manuellement et qui opèrent directement au niveau des tokens. Cette technique, que nous avons développée spécifiquement pour améliorer le traitement des corpus bruités issus du web, permet de développer autour d'un étiqueteur morphosyntaxique comme MElt un *wrappeur* de débruitage selon une approche comparable à celle évoquée à la section 8.1.4 pour l'étiquetage des entités nommées, sans toutefois lui être strictement parallèle. L'idée générale est ici la suivante. La plupart des étiqueteurs morpohsyntaxiques sont adaptés au traitement de corpus édités, souvent journalistiques. Leurs performances sont moindres sur d'autres types de textes, et notamment sur du texte bruité. Il y a donc avantage à débruiter au maximum un texte issu du web avant de l'étiqueter. Toutefois, c'est bien le texte de départ que l'on veut étiqueter, pas une version débruitée. Il faut donc savoir étiqueter le texte de départ à partir des étiquettes obtenues pour la version débruitée du texte.

Cette méthodologie a été utilisée et décrite sur le français dans le cadre du développement du French Social Media Bank (FSMB ; Seddah *et al.*, 2012d) et de l'initiative CoMéRé (Chanier *et al.*, 2014) pour le développement de corpus annotés du français médié par les réseaux, mais également sur l'anglais lors de notre participation à la campagne SANCL 2012 d'évaluation des analyseurs syntaxiques sur des données issues du web (Petrov et McDonald, 2012). Elle peut se résumer en six étapes comme suit.

Pré-traitement : Nous appliquons tout d'abord certaines des grammaires locales de SxPipe qui opèrent au niveau des caractères (smileys, URL, adresses, hashtags Twitter…).

Tokenisation : Le texte brut est tokenisé et segmenté en phrases par le module correspondant d'SxPipe.

Normalisation : La technique décrite à la section 7.3.3 est appliquée. La correspondance entre « tokens corrigés » résultant de cette étape de correction et tokens d'origine est conservée.

Étiquetage : Les tokens corrigés sont étiquetés et lemmatisés par la version standard de MElt, entraînée sur des données éditées [53].

Post-étiquetage : Une quainzaine de règles de post-édition sont appliquées pour améliorer l'étiquetage. Ces règles sont quasiment indépendantes de la langue, et concernent notamment certaines entités spécifiques au web (annotation des URL, des adresses e-mail, des smileys, etc.).

Dénormalisation : Des étiquettes et des lemmes sont attribués aux tokens d'origine à partir des étiquettes et des lemmes obtenus pour les « tokens corrigés », de la façon suivante. Si un token corrigé correspond à un token de départ, son étiquette et son lemme sont attribués à ce dernier. Si un token corrigé correspond à plusieurs tokens de départ, son étiquette et son lemme sont attribués au dernier d'entre eux, tous les autres recevant l'étiquette spéciale *Y* et un lemme vide. Si plusieurs « tokens corrigés » correspondent à un seul token de départ, les différentes étiquettes correspondantes sont concaténées en une *étiquette multiple* à l'aide du séparateur *+*. Il en va de même pour les lemmes. Cette convention est cohérente avec les étiquettes *P+D* et *P+PRO* du FTB, qui correspondent aux amalgames standard du français. Le tableau 8.9 illustre les étiquettes multiples non standard les plus fréquemment rencontrées dans le FSMB.

Illustrons cette méthodologie par quelques exemples. Une séquence comme `l'après midi` sera étiquetée `l'/DET/le après/Y/ midi/NC/après-midi` après découpage en trois tokens et normalisation temporaire par la règle « `l' après`

53. Comme nous le verrons dans la suite de cette section, nous avons utilisé $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ pour le français (cf. chapitre 8) et $\text{MElt}_{\text{en}}^{\text{Onto}}$ pour l'anglais (cf. section 8.4.3).

Étiquette composée	#occurrences	Exemple attesté	Équivalent standard
CLS+V	54	`c`	`c' est`
ADV+CLO	12	`ni`	`n' y`
CS+CLS	12	`qil`	`qu' il`
CLS+CLO	11	`jen`	`j' en`
CLO+V	9	`ma`	`m' a`
DET+NC	9	`lamour`	`l' amour`
ADV+V	7	`non`	`n' ont`

Tableau 8.9 – Étiquettes composées non standard figurant au moins 3 fois dans le French Social Media Bank.

Tokens (1 par ligne)	« Tokens corrigés »	« Tokens corrigés » étiquetés par MElt$_{FR}^{FTB\text{-}UC}$	Tokens étiquetés
i	I	I/*PRP*	i/*PRP*
know	know	know/*VBP*	know/*VBP*
im	I 'm	I/*PRP* 'm/*VBP*	im/*VBP*
gon	going	going/*VBG*	gon/*VBG*
na	to	to/*TO*	na/*TO*
visit	visit	visit/*VB*	visit/*VB*
somewebsite.com	_URL	_URL/*NN*	somewebsite.com/*ADD*

Tableau 8.10 – Illustration du processus d'étiquetage morphosyntaxique de textes bruités par normalisation temporaire sur l'exemple forgé en anglais déjà étudié à la table 7.5.

`␣midi` → `l' après-midi` ». Le token unique `chépa`, quant à lui, est annoté `chépa/CLS+ADV+V+ADV/je+ne+savoir+pas`. Le cas plus complexe de `c t`, déjà évoqué plus haut, produit `c/Y/ t/CLS+V/ce+être` : puisqu'il n'y a pas correspondance entre `c` et `t` d'une part et *c'* et *était* d'autre part, on ne peut que considérer `c t` comme la transcription sous forme de composé de l'amalgame d'un clitique sujet et d'une forme fléchie du verbe être. Les tableaux 8.10 et 8.11 reprennent les exemples des tableaux 7.5 et 7.6 en illustrant cette fois-ci le processus complet qui va jusqu'à l'étiquetage morphosyntaxique final (la lemmatisation, qui est facultative dans MElt, n'est pas indiquée par souci de clarté).

8.4.2 Application au développement d'un corpus arboré de textes bruités issus du web : le cas du French Social Media Bank

Comme indiqué à la section 7.3.3, la première application de MElt$_{fr}^{FTB\text{-}UC}$ avec ce wrappeur de normalisation temporaire fut en tant que pré-annotateur lors du développement du FSMB, et plus spécifiquement son annotation morphosyntaxique. Nous avons alors souhaité poursuivre un double objectif : (i) accélérer l'annotation

Tokens (1 par ligne)	« Tokens corrigés »	« Tokens corrigés » étiquetés par $\text{MElt}_{\text{FR}}^{\text{FTB-UC}}$	Tokens étiquetés	Tokens étiquetés après correction manuelle
sa	ça	ça/*PRO*	sa/*PRO*	sa/*PRO*
fé	fait	fait/*V*	fé/*V*	fé/*V*
o	au	au/*P+D*	o/*P+D*	o/*P+D*
moin	moins	moins/*ADV*	moin/*ADV*	moin/*ADV*
6	6	6/*DET*	6/*DET*	6/*DET*
mois	mois	mois/*NC*	mois/*NC*	mois/*NC*
qe	que	que/*PROREL*	qe/*PROREL*	qe/*CS*
les	les	les/*DET*	les/*DET*	les/*DET*
preliminaires	préliminaires	preliminaires/*NC*	preliminaires/*NC*	preliminaires/*NC*
sont	sont	sont/*V*	sont/*V*	sont/*V*
”	”	”/*PONCT*	”/*PONCT*	”/*PONCT*
sauté	sautés	sauté/*VPP*	sauté/*VPP*	sauté/*VPP*
”	”	”/*PONCT*	”/*PONCT*	”/*PONCT*
c a dire	c’est-à-dire	c’est-à-dire/*CC*	c/*Y* a/*Y* dire/*CC*	c/*Y* a/*Y* dire/*CC*
qil	qu’ il	qu’/*CS* il/*CLS*	qil/*CS+CLS*	qil/*CS+CLS*
yen	y en	y/*CLO* en/*CLO*	yen/*CLO+CLO*	yen/*CLO+CLO*
a	a	a/*V*	a/*V*	a/*V*
presk	presque	presque/*ADV*	presk/*ADV*	presk/*ADV*
pa	pas	pas/*ADV*	pa/*ADV*	pa/*ADV*

Tableau 8.11 – Illustration du processus d’étiquetage morphosyntaxique de textes bruités par normalisation temporaire sur l’exemple français du French Social Media Bank déjà étudié à la table 7.6. Le résultat de l’étape manuelle de finalisation des annotations morphosyntaxiques est également montré.

morphosyntaxique du FSMB en diminuant le temps de travail manuel tout en contribuant à l'homogénéité des annotations, mais également (ii) tester sur quelques sous-corpus de test la pertinence de notre approche. Le premier objectif était justifié entre autres par les résultats que nous avions présentés dans (Fort et Sagot, 2010) selon lesquels une pré-annotation, même en-deçà d'un niveau état-de-l'art, permettait d'accélérer l'annotation manuelle d'un corpus tout en améliorant la qualité de l'annotation.

Le FSMB est un corpus arboré de 1700 phrases en français produites par les utilisateurs de l'internet (Seddah *et al.*, 2012c,d), annoté manuellement en morphosyntaxe et en constituants. C'est à partir du double constat suivant que nous est apparue la nécessité de développer ce corpus. D'une part, la masse de données produites sur internet par tout un chacun et la valeur informationelle de ces données rend de plus en plus utile la capacité à les traiter, et notamment à en faire l'analyse morphosyntaxique et syntaxique. D'autre part, ces données dévient souvent de la norme selon des modalités variables selon les types de données (forums, réseaux sociaux, messagerie instantanée, etc.) mais toujours linguistiquement intéressantes. Ce sont du reste ces raisons qui ont également conduit au développement du Google Web Treebank (Bies *et al.*, 2012), sur lequel nous reviendrons. Les principaux autres corpus de données issues du web alors disponibles étaient les corpus arborés de la Dublin City University couvrant Twitter et des forums de la BBC sur le football (Foster *et al.*, 2011a,b) et le corpus de données Twitter annoté en parties du discours développé par Gimpel *et al.* (2011).

Nous avons décidé de nous concentrer sur deux types de données : (i) les réseaux sociaux, en l'espèce Facebook et Twitter, et (ii) les forums de discussion en ligne, parmi lesquels nous avons retenu un forum médical (Doctissimo [54]) et un forum centré au départ sur les jeux vidéos (JeuxVideo [55]). Ainsi, le FSMB était le premier corpus couvrant des donées issues de Facebook, et le premier corpus de données issues du web pour une langue autre que l'anglais. Des corpus couvrant d'autres types de donnée bruitées existaient déjà, parfois depuis un certain temps, et notamment des corpus de SMS ou de courriers électroniques (notamment au sein du corpus EASy/PASSAGE).

Les données proprement dites ont été choisies de façon à ce que le FSMB reflète au mieux l'éventail des phénomènes non standard et des différents niveaux d'écarts à la norme que l'on rencontre sur internet. À cet égard, le FSMB n'est pas représentatif au sens quantitatif du terme de ce que l'on trouve sur internet, mais il essaye d'être représentatif des phénomènes que l'on y rencontre. Naturellement, une telle représentativité est subjective et critiquable.

Le tableau 8.12 donne quelques informations quantitatives sur les sous-corpus qui constituent le FSMB. Le niveau de bruit y est mesuré par une métrique dédiée, que nous

54. `http://forum.doctissimo.fr`.
55. `http://www.jeuxvideo.com`.

avons développée à partir de la divergence de Kullback-Leibler entre les distributions des caractères dans deux corpus distincts, en prenant pour référence le FTB [56].

	#Phrases	#Tokens	Longueur moyenne	Niveau de bruit
Doctissimo	771	10834	14,05	0,37
B+	36	640	17,78	1,29
B-	735	10 194	13,87	0,31
JeuxVideo	199	3 058	15,37	0,81
Twitter	216	2 465	11,41	1,24
B+	93	1 126	12,11	1,46
B-	123	1 339	10,89	1,08
Facebook	452	4 200	9,29	1,67
B+	120	1 012	8,43	2,44
B-	332	3 188	9,60	1,30

Tableau 8.12 – Propriétés des différents sous-corpus du French Social Media Bank, en fonction notamment de la distinction entre sous-corpus faiblement bruités (B-) et fortement bruités (B+). Les niveaux de bruit sont calculés en prenant le FTB comme référence.

Dans le FSMB, les conventions d'annotation pour le niveau morphosyntaxique sont quasiment les mêmes que celles du FTB-UC (cf. section 8.1.2 ; Candito et Crabbé, 2009), à des fins de compatibilité pour les outils et les évaluations. Quelques adaptations mineures ont été toutefois nécessaires. Tout d'abord, nous utilisons lorsque c'est nécessaire les étiquettes composées non standard, comme illustré précédemment, notamment à la table 8.11. Par ailleurs, nous avons ajouté deux nouvelles étiquettes par rapport au FTB : *HT* pour les hashtags Twitter et *META* pour les tokens méta-textuels tels que `RT` sur Twitter. On notera que les URL, les adresses e-mail et, dans les données Twitter, les *at-mentions* sont toutes étiquetées *NPP* (nom propre). Ceci permet de rester maximalement compatible avec les modèles entraînés sur le FTB, lequel ne contient pas de tokens de cette nature. C'est une différence importante avec d'autres travaux sur les données issues du web, mais ne pose pas de difficulté dans la mesure où identifier un token annoté *NPP* comme étant une URL, une adresse e-mail ou une *at-mention* est trivial.

Nous avons employé deux stratégies de pré-annotation distinctes en fonction du niveau de bruit des différents sous-corpus. Pour les sous-corpus les moins bruités (B- dans la table 8.12), nous avons utilisé une version légèrement modifiée des outils de tokenisation et de segmentation en phrases intégrés à l'architecture Bonsai d'analyse syntaxique, qui

56. Cela ne signifie pas que nous considérons le FTB comme représentatif du français standard. Nous utilisons le FTB comme référence pour mesurer le niveau de bruit en tant que c'est sur ce corpus que sont entraînées les versions standard des outils comme $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ ou l'analyseur syntaxique que nous utiliserons au chapitre suivant sur ces mêmes données. Ainsi, l'écart au FTB est une mesure pertinente pour estimer le degré d'inadéquation de tels outils sur les données considérées.

repose sur le FTB-UC (Candito *et al.*, 2010), suivie de la version standard de MORFETTE [57]. Cela nous a permis de rester aussi proches que possible des conventions de tokenisation, de segmentation et d'étiquetage morphosyntaxique du FTB-UC.

Pour les sous-corpus les plus bruités (B+ dans la table 8.12), nous avons utilisé $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ avec le wrappeur de normalisation qui fait l'objet de cette section.

Nous avons toutefois évalué l'apport du wrappeur de normalisation temporaire sur la qualité de l'annotation morphosyntaxique produite par $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ sur les deux types de sous-corpus [58]. De plus, nous avons réservé certains sous-corpus comme corpus de test. Ces sous-corpus n'ont donc pas été utilisés pour le développement du wrappeur : ils n'ont pas été utilisés pour l'extraction de tokens et de séquences de tokens candidats à la construction d'une règle de normalisation (cf. section 8.4.1). Nous avons donc produit des résultats séparés sur ces sous-corpus de test et sur les autres, qui jouent donc le rôle de corpus de développement.

Les résultats de ces évaluations sont fournis à la table 8.13. Ils montrent que l'utilisation du wrappeur de normalisation temporaire améliore les résultats de façon importante, y compris sur les corpus de test, et ce sans pour autant dégrader ceux obtenus sur le FTB-TEST. On notera que nos résultats sur les sous-corpus Twitter sont similaires à ceux obtenus par Foster *et al.* (2011a) sur des données Twitter en anglais, quand bien même ces résultats ne sont pas directement comparables puisqu'ils concernent des langues et des jeux d'étiquettes différents. Par ailleurs, on peut observer que les scores obtenus sont corrélés au niveau de bruit tel que mesure par notre métrique, et ce que l'on utilise ou non le wrappeur de normalisation temporaire [59].

8.4.3 Expériences sur le Google Web Treebank

Lors du développement du FSMB évoqué à la section précédente, MElt, couplé au wrappeur de normalisation temporaire destiné aux textes bruités, a été utilisé sur des données en grande partie connues à l'avance. Certes, nous avons fait attention à mettre de côté des sous-corpus de test afin d'évaluer la qualité de nos traitements, mais le contexte n'en était pas moins celui d'une pré-annotation et pas d'une annotation de textes pleinement inconnus. À l'inverse, nos expériences sur l'anglais ont été réalisées dans le cadre de notre participation (Seddah *et al.*, 2012b) à la campagne SANCL 2012 d'évaluation

57. La raison pour laquelle MORFETTE a été utilisé à la place de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ est double : d'une part l'étiquetage produit par MORFETTE est quasiment au même niveau que celui de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$, et d'autre part la lemmatisation n'était alors pas encore réalisée par $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ de façon suffisamment satisfaisante.

58. La version de $\text{MElt}_{\text{fr}}^{\text{FTB-UC}}$ utilisée est une version légèrement améliorée de celle décrite plus haut. Ces modifications mineures expliquent l'écart, faible, entre le score de 97,75% mentionné au chapitre 8 et le score de 97,79% mentionné à la table 8.13 sur le FTB-TEST.

59. Des régressions linéaires simples donnent des pentes de -4,8 et -7,2 selon que l'on utilise ou non le wrappeur, avec des coefficients de corrélation de respectivement 0,77 et 0,88.

	Corpus de développement		Corpus de test	
	$MElt_{fr}^{FTB\text{-}UC}$	$MElt_{fr}^{FTB\text{-}UC}$+corr	$MElt_{fr}^{FTB\text{-}UC}$	$MElt_{fr}^{FTB\text{-}UC}$+corr
Doctissimo				
B+	56,41	80,78	–	–
B-	86,57	88,42	87,78	89,18
JeuxVideo	81,20	82,41	82,64	83,63
Twitter				
B+	80,21	84,51	74,50	81,65
B-	84,09	89,00	86,23	88,24
Facebook				
B+	–	–	67,00	70,75
B-	71,75	76,87	78,66	82,00
Total	*80,64*	*84,72*	*83,10*	*85,28*
FTB	97,42	97,42	97,79	97,78

Tableau 8.13 – Résultats pour l'étiquetage morphosyntaxique de chaque sous-corpus du French Social Media Bank, en fonction notamment de la distinction entre sous-corpus faiblement bruités (B-) et fortement bruités (B+). Les colonnes $MElt_{fr}^{FTB\text{-}UC}$ donnent les résultats de la version standard de MElt (version de 2012), sans application de l'outil de débruitage mais avec utilisation des grammaires locales de base (URL, e-mails, smileys). Les colonnes $MElt_{fr}^{FTB\text{-}UC}$+corr donnent les résultats obtenus en rajoutant le débruitage.

des analyseurs syntaxique de l'anglais sur des données issues du web (Petrov et McDonald, 2012). Les organisateurs avaient fourni aux participants les données suivantes :

— Données annotées en morphosyntaxe et en constituants : la version Ontonotes 4.0 du Penn TreeBank (sous-corpus d'entraînement et de développement) et les sous-corpus d'e-mails et de blogs du Google Web TreeBank (Bies *et al.*, 2012) en tant que données de développement (mais les sous-corpus de textes issus de newsgroups, des questions posées sur *Yahoo! answers* et des recensions [60], qui ont servi de sous-corpus de test).

— Données brutes : 70 millions de tokens de texte brut couvrant de façon inéquitable ces cinq domaines [61]

La méthodologie décrite dans cette section, appliquée avec les règles de correction évoquées et illustrées à la section 7.3.3, nous ont permis d'obtenir sur les données de développement les résultats présentés dans le tableau 8.14. On y constate que les meilleurs résultats sont généralement obtenus en couplant MElt avec application des règles de débruitage. Nous reviendrons à la section 9.4.2 sur les résultats obtenus par notre équipe

60. Nous traduisons par *recension* l'anglais *review*.
61. De 27 000 à 2 millions de phrases selon les domaines.

en analyse syntaxique grâce à cette stratégie d'étiquetage morphosyntaxique, qui nous ont permis d'être classés deuxièmes.

	Tous les mots			Mots inconnus		
	Morfette	$\text{MElt}^{\text{Onto}}_{\text{en}}$	$\text{MElt}^{\text{Onto}}_{\text{en}}$+corr	Morfette	$\text{MElt}^{\text{Onto}}_{\text{en}}$	$\text{MElt}^{\text{Onto}}_{\text{en}}$+corr
E-mails	88,8	88,9	**90,4**	61,3	62,4	**72,1**
Blogs	93,9	**94,7**	**94,7**	78,7	87,2	**87,3**
Ontonotes (dev)	96,4	**96,5**	**96,5**	**93,2**	92,3	92,9

Tableau 8.14 – Résultats pour l'étiquetage morphosyntaxique de chaque sous-corpus du Google Web Treebank fourni comme corpus de développement lors de la campagne SANCL 2012. Les colonnes $\text{MElt}^{\text{Onto}}_{\text{en}}$ donnent les résultats de la version standard de MElt entraînée sur la section d'entraînement d'Ontonotes, sans application de l'outil de débruitage mais avec utilisation des grammaires locales de base (URL, e-mails, smileys). Les colonnes $\text{MElt}^{\text{Onto}}_{\text{en}}$+corr donnent les résultats obtenus en rajoutant le débruitage. Les résultats sont comparés avec ceux de Morfette (Chrupała *et al.*, 2008).

8.5 Éléments de conclusion

Les expériences décrites dans ce chapitre montrent toutes la pertinence, à vrai dire peu surprenante, de l'utilisation d'informations lexicales issues d'un lexique morphosyntaxique pour améliorer les performances d'un étiqueteur morphosyntaxique, qu'il s'appuie sur une architecture statistique ou sur une architecture neuronale. Elles montrent également l'importance de la prise en compte du caractère bruité des corpus non-standard, et notamment ceux issus du web.

Concernant la première de ces conclusions, nous ne répéterons pas ici les remarques faites à la section 8.3.4. Il nous suffira d'indiquer que les approches neuronales n'ont pas, à ce stade, montré une supériorité massive sur les approches statistiques en étiquetage morphosyntaxique, notamment lorsque l'on ne dispose que d'un volume restreint de données d'apprentissage, et *a fortiori* lorsque l'on n'a pas de gros volumes de textes bruts pour pré-entraîner des w-embeddings. La raison en est peut-être que la tâche d'étiquetage en parties du discours est suffisamment bien comprise pour qu'un inventaire de classes de traits conçu manuellement puisse faire aussi bien, voire mieux, qu'un réseau de neurones dont l'un des avantages est de choisir les traits pertinents de façon automatique. Il n'en reste pas moins que les réseaux de neurones continuent à progresser, avec des architectures nouvelles qui pourraient apporter des gains plus importants et mieux gérer le cas des corpus d'entrainement de taille réduite. Ainsi, Inoue *et al.* (2017) décrivent et évaluent sur des données arabes une architecture très similaire à celle d'alNNtagger pour

l'intégration à un réseau de neurone récurrent d'informations issues d'un lexique externe, mais en couplant la prédiction des étiquettes UPOS et celle des traits morphosyntaxiques, à l'image de ce que nous avons fait dans un cadre statistique avec alVWtagger.

Nos expériences sur l'étiquetage de corpus bruités méritent en revanche un regard plus critique. En effet, l'approche par règles que nous avons développé pour la normalisation temporaire de tels corpus en vue de leur étiquetage s'est révélée moins générale qu'il n'y paraissait de prime abord, au vu des résultats présentés à la section précédente. En réalité, la variation linguistique est très élevée dans ce type de données, et cette variation elle-même évolue au fil des mois et des années. C'est donc bien naturellement que des techniques plus adaptatives qu'un simple jeu de règles ont été développées récemment. De façon plus intéressante, Gui *et al.* (2017) présentent une architecture neuronale exploitant la notion d'apprentissage adversarial et différentes sources de données (y compris hors-domaine et/ou non annotées) pour l'étiquetage de données issues de Twitter, avec de très bons résulats. C'est certainement vers ce type de travaux, ou à tout le moins vers le développement de techniques non ou faiblement supervisées de normalisation de textes bruités, qu'il faut chercher des solutions au traitement de corpus bruités, à commencer par leur étiquetage morphosyntaxique.

Chapitre 9

Analyse syntaxique et informations lexicales

Sommaire

Le chapitre précédent a montré l'impact positif des informations lexicales morphologiques pour l'étiquetage morphosyntaxique. Dans ce chapitre, nous allons nous pencher sur le rôle que peuvent jouer des informations lexicales sur l'analyse syntaxique. Pour un aperçu de l'historique de la recherche en analyse syntaxique, on pourra se reporter à la section A.10.

Les corpus arborés ne sont pas suffisamment grands et les modèles probabilistes de la syntaxe pas véritablement adaptés pour extraire et exploiter en tant que telles des informations lexicales riches telles que les informations lexico-syntaxiques, et notamment les informations de sous-catégorisation sur les unités prédicatives : ceci vient des problèmes de dispersion des données ainsi que de la localité trop restreinte des modèles. À cet égard, l'avènement des modèles neuronaux ne change que partiellement la donne : si la problématique de la dispersion des données peut être partiellement réduite par l'utilisation d'embeddings et la non-linéarité par celle de modèles comme les LSTM capables de faire transiter de l'information d'un endroit à l'autre de la phrase, les informations de sous-catégorisation ne sont pas modélisées en tant que telles. Or ces informations sont clairement cruciales pour l'analyse sytaxique : on peut s'en convaincre par exemple d'une part au travers du rôle central qu'elles jouent en analyse syntaxique symbolique (cf. section 9.1), et d'autre part à la vue de certains types d'erreurs effectués par les analyseurs statistiques en constituants, fussent-ils neuronaux, au niveau des structures argumentales. La raison fondamentale en est qu'un lexique syntaxique est une collection d'informations linguistiques de nature différente et complémentaire de celles qui constituent un corpus arboré (ou une grammaire). Un tel lexique contient en effet, comme nous l'avons vu notamment au chapitre 5, des informations renseignées manuellement, parfois induites automatiquement, dont la majorité ne peuvent être extraites directement de corpus arborés, en raison du problème de la dispersion des données. On peut tenter de contourner ce problème au moyen de statistiques distributionnelles sur des corpus bruts de taille gigantesque [1], mais le recours à des lexiques syntaxiques riches reste une solution raisonnable, pour au moins trois raisons : (i) de telles ressources existent au moins pour certaines langues, (ii) développer (manuellement ou semi-automatiquement) de telles ressources n'a rien de moins « noble » que développer (manuellement ou semi-automatiquement) des corpus arborés, et (iii) le recours à des corpus bruts très volumineux n'est pas très satisfaisant : cela nécessite des

1. Dont l'ordre de grandeur est le milliard de mots ou au-delà.

moyens computationnels importants, il n'est pas clair qu'on apprenne grand chose sur la langue en procédant de la sorte, et surtout, intuitivement, le recours à des informations distributionnelles à grande échelle est presque explicitement un moyen de calculer des *approximations* d'informations intrinsèquement plus pertinentes, parmi lesquelles les informations lexico-syntaxiques [2].

La prise en compte dans les analyseurs statistiques d'informations lexicales de niveau syntaxique (informations de valence, notamment), et pas seulement morphologique, est cependant une problématique complexe et relativement peu étudiée. Il s'agit en apparence d'un paradoxe, dans la mesure où les travaux de modélisation syntaxique des dernières décennies du XX^{e} siècle avaient au contraire mis l'accent sur le rôle central de la sous-catégorisation, comme rappelé à la section A.5. C'est notamment le cas de formalismes comme les grammaires lexicales fonctionnelles (LFG), les grammaires syntagmatiques guidées par les têtes (HPSG) ou les grammaires d'adjonction d'arbres (TAG), qui intègrent directement les informations de sous-catégorisation dans les éléments de base des grammaires. Ainsi, comme indiqué plus haut, ce type d'informations est directement exploité par les analyseurs symboliques reposant sur de tels formalismes. La difficulté d'intégrer les informations lexico-syntaxiques dans des analyseurs statistiques tient en ceci que ces derniers reposent souvent soit sur des modèles génératifs structurellement simples (la complexité est modélisée par les informations statistiques) soit sur des modèles discriminants qui prennent leurs décisions de façon trop locale pour que les informations de sous-catégorisation puissent être exploitées facilement.

C'est Collins (1997) qui a ouvert la voie à l'utilisation d'informations lexico-syntaxiques par un analyseur syntaxique statistique. Il a proposé trois modèles génératifs dont le troisième, le modèle 2, intègre des informations de sous-catégorisation : il calcule tout d'abord les probabilités des cadres de sous-catégorisation possibles puis ajoute comme contrainte le cadre le mieux placé. Cette approche a été appliquée par la suite au français par Arun et Keller (2005). Du côté des approches en dépendances, on peut citer le travail de Zeman (2002), qui a montré que l'utilisation d'informations de fréquences sur les cadres de sous-catégorisation permet d'améliorer les performances de l'analyseur. Un autre travail significatif est celui de Versley et Rehbein (2009), qui a montré sur l'allemand que l'on obtenait une amélioration significative à partir d'un analyseur génératif probabiliste classique, de type PCFG, par l'enrichissement des non-terminaux du corpus arboré au moyen d'informations morphologiques et syntaxiques, quoique dans des proportions limitées, couplé à un modèle discriminant pour choisir la meilleure analyse parmi les arbres de probabilité supérieure à un certain seuil. Mais la prise en compte de toute la

2. Mais parmi lesquelles également des informations qui relèvent de ce que l'on appelle parfois la connaissance du monde, et qui correspond plutôt à la représentation cognitive que les locuteurs se font du monde dont ils parlent. De telles informations n'ont rien de linguistique, mais sont détectables dans les corpus en tant qu'elles dessinent les contours de ce que l'on peut vouloir dire ou écrire.

richesse d'une ressource lexico-syntaxique comme le Le*fff* dans un analyseur statistique reste aujourd'hui une question ouverte.

Dans la suite de ce chapitre, nous illustrerons les problématiques à l'interface entre lexique et analyse syntaxique au moyen de quelques études de cas complémentaires, qui concernent, sur différentes langues, l'analyse symbolique et l'analyse statistique, l'analyse en constituants et l'analyse en dépendances, l'intégration d'informations morphologiques et syntaxiques, et le traitement de données éditées comme bruitées :

— À la section 9.1, nous nous concentrons sur l'analyse statistique symbolique. Nous étudions l'intégration du Le*fff* au sein de l'analyseur symbolique FRMG mentionné précédemment, qui repose sur le formalisme des TAG et sur les informations morphologiques et syntaxiques fournies par le Le*fff*. Nous comparerons les résultats de FRMG selon qu'il utilise le Le*fff* ou d'autres ressources lexico-syntaxiques.

— La section 9.2 est consacrée à la prise en compte d'informations lexicales morphologiques dans un analyseur statistique en constituants. Nous montrerons, sur le cas de l'analyse syntaxique de l'espagnol, que les informations morphologiques issues du lexique Alexina Le*ff*e et intégrées à l'étiqueteur morphosyntaxique MElt ou à un lemmatiseur permettent d'améliorer les performances d'un analyseur à la Petrov *et al.* (2006) jusqu'à atteindre des performances état-de-l'art, surtout si l'on choisit le bon niveau de granularité pour le jeu d'étiquettes morphosyntaxiques.

— Nous verrons à la section 9.3 comment les informations syntaxiques du Le*fff* peuvent être utilisées pour améliorer un analyseur syntaxique statistique en dépendances du français, en les comparant à d'autres types d'informations syntaxiques que l'on peut extraire du corpus d'entraînement lui-même ou d'un corpus préalablement analysé automatiquement.

— Nous évoquerons enfin à la section 9.4 nos travaux sur l'analyse syntaxique de corpus bruités (notamment issus du web), qui font suite aux travaux en analyse morphosyntaxique sur les mêmes données décrits à la section 8.4.

9.1 Informations morphologiques et syntaxiques pour l'analyse syntaxique symbolique : comparaison du Le*fff* avec d'autres lexiques [3]

Comme évoqué précédemment, les analyseurs syntaxiques symboliques reposent souvent sur des informations lexicales riches, et notamment syntaxiques. Nous avons d'ailleurs déjà mentionné, notamment au chapitre 5, le fait que le Le*fff* ait été intégré à des analyseurs syntaxiques symboliques du français reposant sur des formalismes variés (cf. section 5.1.2), grâce à son modèle lexico-syntaxique qui ne dépend pas d'un formalisme syntaxique particulier.

Nous reviendrons sous peu sur les caractéristiques de l'analyseur symbolique FRMG, sur lequel reposent les expériences décrites dans cette section. Le fait qu'il repose sur une grammaire d'adjonction d'arbres lexicalisée couplée aux informations lexicales morphologiques et syntaxiques fournies par le Le*fff* permet d'étudier l'impact de ces informations sur la qualité des analyses produites. En particulier, on peut envisager de remplacer en tout ou partie les informations extraites du Le*fff* par des informations lexicales issues d'autres lexiques, afin de comparer leur pertinence pour l'analyse syntaxique. Cette section décrit des expériences de cette nature, au cours desquelles nous avons comparé les résultats de FRMG selon qu'il utilise le Le*fff* ou que les informations syntaxiques qu'il contient sont remplacées, pour les seules entrées verbales, par des informations syntaxiques fournies par d'autres ressources lexicales dont nous avons déjà parlé aux chapitres précédents, à savoir les tables du Lexique-Grammaire et le lexique DICOVALENCE (Tolone *et al.*, 2011, 2012). Une telle comparaison nécessite naturellement de convertir ces deux ressources en lexiques Alexina, et plus précisément en lexiques faisant usage des mêmes inventaires de fonctions syntaxiques, réalisations, redistributions et autres informations syntaxiques que le Le*fff*. Ces versions converties sont appelées respectivement LG*Lex*$_{\text{Le}fff}$ et DICOVALENCE$_{\text{Le}fff}$ dans la suite de ce chapitre. Concernant le Le*fff*, nous nous sommes appuyés pour ces expériences sur sa version 3.0 mais aussi sur une version alternative dans laquelle les entrées verbales ont été remplacées par le résultat de la fusion avec DICOVALENCE et après un travail de validation manuelle, comme évoqué à la section 5.2.1. Cette version alternative est appelée NewLe*fff* dans la suite de cette section.

La table 9.1 fournit quelques données quantitatives élémentaires sur les quatre lexiques Alexina utilisés, les deux lexiques obtenus par conversion et les deux versions du Le*fff*.

3. Travail réalisé en collaboration avec Éric Villemonte de La Clergerie (ALPAGE) et Elsa Tolone (alors doctorante à l'Université Paris-Est Marne-la-Vallée, au sein de l'équipe d'informatique linguistique du LIGM). Ces travaux ont fait l'objet de plusieurs publications (Sagot et Tolone, 2009a,b ; Tolone et Sagot, 2009, 2011 ; Tolone *et al.*, 2011, 2012).

Lexique	#Entrées	#Lemmes	#Entrées/#Lemmes
Le*fff*	7 108	6 827	1,04
LG*Lex*$_{\text{Le}fff}$	13 867	5 738	2,41
Dicovalence$_{\text{Le}fff}$	8 313	3 738	2,22
NewLe*fff*	12 613	7 933	1,58

Tableau 9.1 – Données quantitatives sur les quatre lexiques syntaxiques verbaux utilisés

9.1.1 L'analyseur syntaxique frmg

9.1.1.1 Description

frmg (FRench MetaGrammar) est un analyseur syntaxique profond à large couverture pour le français [4] (Villemonte de La Clergerie, 2005 ; Thomasset et Villemonte de La Clergerie, 2005 ; Villemonte de La Clergerie *et al.*, 2009a ; Villemonte de La Clergerie, 2013, 2014). En réalité, l'acronyme frmg désigne des ressources couvrant plusieurs niveaux de représentation des informations linguistiques. Le niveau le plus abstrait est celui d'une métagrammaire modulaire, hiérarchique et linguistiquement motivée. Cette métagrammaire permet de produire une grammaire d'adjonction d'arbres (TAG) compacte et faisant usage de structures de traits. Elle contient environ 300 arbres élémentaires factorisés, dont 35 peuvent être ancrés par un verbe. En dépit de sa compacité, cette grammaire a une couverture importante grâce à divers opérateurs de factorisation, tels que des disjonctions et des contraintes (ou *gardes*), qui permettent d'autoriser plusieurs parcours différents d'un même arbre.

Cette grammaire est elle-même compilée en un analyseur syntaxique tabulaire efficace, lui aussi nommé frmg, qui est capable de produire soit des analyses complètes, dès lors que c'est possible, soit des analyses partielles, toutes représentées sous forme de forêts partagées de dépendances. Ces forêts peuvent ensuite être désambiguïsées à l'aide de règles heuristiques pondérées (manuellement ou par apprentissage supervisé) afin d'en extraire des arbres de dépendances uniques pour chaque phrase. Enfin, ces arbres peuvent être ensuite convertis dans différents formalismes et formats, y compris celui des campagnes d'évaluation EASy/PASSAGE (Paroubek *et al.*, 2006 ; Hamon *et al.*, 2008 ; Paroubek *et al.*, 2009) ou le format des campagnes CoNLL (Nivre *et al.*, 2007a).

frmg bénéficie du domaine de localité étendu des grammaires d'adjonction d'arbres. Ceci permet par exemple de capturer tous les arguments prenant part à un même cadre de sous-catégorisation au moyen d'un unique arbre élémentaire. Mais cela nécessite naturellement que frmg dispose d'informations lexico-syntaxiques riches, ce qui est le cas grâce au Le*fff*. Concrètement, chaque arbre élémentaire de frmg est associé à un *hypertag* (Kinyon, 2000), une structure de traits qui, dans le cas d'une unité

4. frmg est librement disponible à l'adresse `http://mgkit.gforge.inria.fr/`

prédicative, résume notamment les différentes sous-catégorisations possibles (fonctions syntaxiques et réalisations) couvertes par l'arbre. En parallèle, les entrées du Le*fff* sont également converties en *hypertags*. L'ancrage d'un arbre par une entrée lexicale nécessite l'unification des deux hypertags correspondants.

9.1.1.2 Performances et comparaison avec l'état de l'art

La section A.10.3 propose un aperçu de l'état de l'art de l'analyse syntaxique du français fin 2014, dont nous avons reproduit le tableau A.4 (ici 9.2) qui compare les performances de différents analyseurs syntaxiques avec la métrique LAS telles que mesurées sur la section de test FTB-TEST du Corpus Arboré de Paris 7. Il permet de situer FRMG au sein de l'écosystème des analyseurs syntaxiques du français, toutes approches confondues [5]. La quasi totalité de ces analyseurs syntaxiques utilisent le Le*fff*, soit au sein de MElt, soit en en exploitant les informations morphologiques et parfois syntaxiques. On constate qu'un analyseur symbolique comme FRMG obtient des performances tout à fait honorables. De plus, Villemonte de La Clergerie (2013, 2014) montre que les performances de FRMG se dégradent moins que celles de l'analyseur de Berkeley lorsque l'on passe d'un corpus journalistique à un corpus médical. Il est donc très intéressant de constater que, contrairement à ce qui a pu être estimé par moments, les techniques d'analyse syntaxique reposant sur des grammaires génératives et des algorithmes performants ont toujours toute leur place, notamment lorsqu'elles sont couplées à des techniques de désambiguïsation statistiques voire neuronales. Il est important de rappeler qu'un analyseur statistique ou neuronal ne peut être construit qu'à partir d'un corpus arboré, corpus dont le développement a un coût considérable. Le coût de développement des ressources linguistiques que sont une grammaire comme FRMG et un lexique comme le Le*fff* ne sauraient donc être comparées au seul temps d'entraînement d'un analyseur syntaxique statistique ou neuronal, mais doivent l'être au temps de développement du corpus arboré sous-jacent. Naturellement, les analyseurs statistiques et neuronaux ont des performances et une robustesse indéniables. De plus, l'annotation de corpus arborés peut être confiée à des annotateurs moins pointus dans leur expertise que s'il leur fallait développer une métagrammaire comme celle de FRMG. C'est donc dans l'hybridation entre techniques symboliques, statistiques et désormais neuronales que réside certainement la voie la plus prometteuse. C'est du reste dans cet ordre d'idées que FRMG a été couplé par Villemonte de La Clergerie (2014) avec un analyseur statistique à base de transitions développé à cette fin, DyALog-SR, conduisant, toujours avec le Le*fff* comme ressource lexicale, aux meilleurs scores LAS jamais publiés sur l'analyse syntaxique du français sur le FTB, à savoir 90,25%.

5. La suite de ce paragraphe reprend le dernier paragraphe de la section A.10.3.

Analyseur	Références	LAS (%)
Berkeley	Petrov *et al.* (2006) ; Petrov et Klein (2007) ; Candito *et al.* (2010)	86,80
frmg	Villemonte de La Clergerie (2013) ; cf. section 9.1	87,17
MaltParser	Nivre *et al.* (2007b) ; Candito *et al.* (2010)	87,30
MSTParser	McDonald *et al.* (2005) ; Candito *et al.* (2010)	88,20
Talismane	Urieli et Tanguy (2013)	88,50
LORG + reranking	Le Roux *et al.* (2012b)	89,00
MElt +DyALog-SR	Villemonte de La Clergerie (2014)	89,01
MElt + LORG + reranking	Le Roux *et al.* (2012b)	89,20
MElt + MATE	Bohnet (2010) ; Le Roux *et al.* (2012b)	89,20
(MElt + DyALog-SR) + frmg	Villemonte de La Clergerie (2014)	90,25

Tableau 9.2 – Scores LAS de différents analyseurs syntaxiques du français sur la version en dépendances du FTB (entraînement sur FTB-train et évaluation sur FTB-test ; ces sous-corpus sont définis comme au chapitre précédent)

9.1.2 Évaluation comparative des résultats obtenus avec les différents lexiques

Nous avons effectué deux séries d'évaluation :

— Une première sur les données et avec les métriques des campagnes EASy/PASSAGE mentionnées plus haut ; nous renvoyons notamment à (Tolone *et al.*, 2011, 2012) pour la description et la discussion des résultats de cette première série d'évaluations.

— Une seconde sur la section d'évaluation du FTB tel que converti par Candito *et al.* (2010) en dépendances, et représenté dans le format CoNLL, format largement utilisé dans les campagnes internationales d'évaluation des analyseurs syntaxiques (Nivre *et al.*, 2007a).

La table 9.3, où l'on rappele également les résultats obtenus avec le Le*fff* par des versions ultérieures de frmg (utilisant le Le*fff*), montre que tous les lexiques ont une très bonne couverture sur le FTB, corpus de style journalistique — la couverture est ici le taux de phrases recevant une analyse complète (toutes les autres, ou presque, recevant des analyses partielles). On constate par ailleurs des performances légèrement meilleures avec le Le*fff* qu'avec les autres lexiques en termes de LAS. Enfin, le fait que LG*Lex*$_{\text{Le}fff}$ soit à la fois à large couverture et à granularité très fine a un impact important sur les temps d'analyse.

Une analyse plus fine en termes de rappel et de précision pour chacun des types de dépendances montre des résultats contrastés, LG*Lex*$_{\text{Le}fff}$ ou plus spécialement NewLe*fff* étant parfois supérieurs soit en rappel, par exemple pour les clitiques verbaux non-sujet les objet indirect en *à*, soit en précision, comme par exemple pour les attributs de l'objet ou les auxiliaires de causatif. On identifie également des limites de certaines ressources, notamment le manque de couverture de Dicovalence$_{\text{Le}fff}$ et de LG*Lex*$_{\text{Le}fff}$ sur

les attributs (du sujet comme de l'objet). De façon plus générale, le rappel est relativement bon mais la précision pose parfois problème. Nous suspectons que la granularité plus fine de LG*Lex*$_{\text{Le}\textit{fff}}$ et, dans une moindre mesure, de NewLe*fff* tend à conduire l'analyseur à sélectionner des entrées lexicales rares pour certains verbes de fréquence moyenne ou élevée, et ce en raison des caractéristiques des heuristiques de désambiguïsation utilisées par FRMG. Des heuristiques telles que celle consistant à préférer interpréter un complément comme un argument plutôt qu'un modifieur lorsque les deux sont possibles peuvent poser des problèmes lorsque le lexique mentionne comme possible des arguments rares ou des arguments que le FTB considère comme des modifieurs. Il en résulte des confusions entre des dépendances argumentales (objets directs et indirects) et des dépendances indiquant un modifieur.

LEXIQUE	COUVERTURE (%)	LAS (%)	TEMPS D'ANALYSE (S)
Le*fff*	89,53	82,21	0,61
NewLe*fff*	88,76	81,36	0,94
LG*Lex*$_{\text{Le}\textit{fff}}$	86,73	78,75	1,95
DICOVALENCE$_{\text{Le}\textit{fff}}$	75,28	79,38	0,69
Le*fff* (FRMG 2013)	–	87,17	–
Le*fff* (FRMG + DYALOG-SR)	–	90,25	–

TABLEAU 9.3 – Évaluation sur la section de test du FTB en dépendances au format CoNLL

9.1.3 Fouille d'erreurs

Ces évaluations fournissent déjà un bon aperçu des qualités relatives des différents lexiques. Cependant, afin d'extraire des informations pertinentes au niveau des entrées verbales individuelles, nous avons repris et adapté la technique de fouille d'erreurs décrite à la section 5.2.2. Nous aurions pu l'utiliser directement, pour procéder sur les lexiques autres que le Le*fff* de la même façon que nous avons procédé pour le Le*fff* à cette même section. Mais il nous a semblé plus intéressant d'adapter la méthode pour permettre de comparer les autres lexiques au Le*fff* de façon contrastive, minimisant ainsi les problèmes liés à la grammaire FRMG. Nous avons procédé de la façon suivante : plutôt que de chercher les mots suspects parmi les phrases inanalysables, nous avons cherché les *verbes* suspects parmi les phrases analysables avec le Le*fff* mais inanalysables avec le lexique à lui comparer. Autrement dit, un verbe donné est d'autant plus suspect pour un lexique alternatif qu'il a tendance à être présent plus qu'attendu dans des phrases analysables avec le Le*fff* mais pas avec ce lexique alternatif.

Cet algorithme modifié a été appliqué sur le résultat de l'analyse d'un corpus de 1,6 million de mots (100 000 phrases), le Corpus Passage Jouet, développé dans le cadre du projet ANR PASSAGE. Ce corpus rassemble des documents de différents genres

(encyclopédique issu de la wikipédia française, littéraire issu de wikisources, dépêches AFP, discours parlementaire issu d'Europarl).

À titre d'exemple, l'analyse des résultats pour LG*Lex*$_{\text{Le}fff}$ a permis d'identifier plusieurs classes de problèmes dans cette ressource :

- des entrées qui sont absentes des tables, pourtant supposées très couvrantes (*mixer* au sens musical, *zapper* 'oublier') ou pour lesquelles certaines propriétés dérivationnelles valides sont codées comme invalides, comme la préfixation en *re-* (propriété *re-V*) sur *affirmer* pour *réaffirmer* et sur *élire* pour *réélire*, erreur parfois cumulée avec un codage erroné de la capacité à prendre part à une construction pronominale (ainsi, *implanter* n'a aucune de ces deux propriétés, d'où des erreurs avec *se réimplanter*) ;
- des entrées mal ou pas encore codées (*susciter*, *réprouver*) ;
- des entrées pourvues d'arguments codés comme obligatoires, alors qu'ils sont attestés sans ces compléments dans le corpus (*délocaliser (Obj)*, *rediriger (Obj)*, *kidnapper Obl (Oblà)*, *revendre Obl (Oblà)*, *écrouer Obl (Loc)*, *camper (Loc)*[6] ;
- des cas particuliers, comme le fait que l'objet direct de *consoler* ne soit pas codé comme cliticisable, ce qui est erroné [7].

9.1.4 Discussion

Les résultats de ces expériences montrent qu'un analyseur symbolique comme FRMG ne tire pas nécessairement parti d'une ressource à la couverture et à la granularité trop fine, comme peut l'être LG*Lex*. La raison en est que les techniques heuristiques utilisées par FRMG pour désambiguïser la forêt d'analyses qu'il construit pour une phrase reposent sur des hypothèses de plausibilité des analyses fournies. Une heuristique fort répandue consiste par exemple à préférer les arguments aux modifieurs. Si, comme dans LG*Lex*, un verbe comme *affamer* est codé comme ditransitif, au titre de la possibilité douteuse de construire des énoncés comme *Marie affame Luc de viande*, l'étape de désambiguisation aura tendance à identifier dès que possible un objet indirect en *de* attachés à *affamer*, y

6. Pour *camper* la situation est la suivante. L'entrée de *camper* qui est en cause est codée dans la table 38LRH, table de verbes à objet direct et à argument supplémentaire locatif (*Le général campe ses troupes dans la plaine*), avec possibilité d'effacer le sujet et de mettre l'objet en position sujet (*Les troupes campent dans la plaine*). Or, dans ce dernier cas, le complément locatif est codé comme obligatoire, ce qui est inexact.

7. Voici également quelques exemples concernant la version de NewLe*fff* alors disponible. Le verbe classé comme étant le plus suspect, et de loin, était *estimer*. Sur 569 phrases contenant une forme de ce verbe, pas moins de 200 étaient inanalysables avec NewLe*fff*. Un rapide coup d'œil à quelques unes de ces phrases a permis d'identifier et de corriger très rapidement l'erreur : l'entrée de l'un des trois lexèmes de forme de citation *estimer*, celle signifiant *considérer (que P)*, n'autorisait pas les réalisations phrastiques de son objet direct (infinitive et complétive). D'autres erreurs : l'information de contrôle manquait à l'entrée pour *s'attendre (à)*, il manquait la réalisation infinitive pour le sujet de *inciter (à)*, il manquait les réalisations phrastiques pour l'objet direct de *réitérer* 'répéter', il manquait les réalisations clitique et complétive pour l'objet indirect de *se résoudre (à)*, et bien d'autres.

compris lorsque ce n'est pas pertinent, c'est-à-dire dans (presque) tous les cas (cf. *La guerre affame le sud du Darfour*, où *du Darfour* ne doit pas être considéré comme un objet indirect de *affame*).

Un autre enseignement est qu'un lexique n'est véritablement utile, dans un contexte comme celui de FRMG, que dès lors qu'il couvre correctement tous les phénomènes que l'analyseur s'attend à y trouver. Par exemple, la mauvaise couverture des arguments attributifs par LG*Lex* et DICOVALENCE pénalise ces ressources.

Enfin, le fait que FRMG et le L*efff* aient été développés en parallèle est inévitablement de nature à favoriser ce dernier par rapport à d'autres lexiques, la grammaire et les heuristiques de désambiguïsation et le contenu du L*efff* ayant été adaptés les uns aux autres. De plus, les nombreuses expériences d'analyse syntaxique de gros corpus avec FRMG (utilisant le L*efff*) qui ont été réalisées depuis des années ont conduit à identifier des erreurs et des manques dans le L*efff* (cf. section 5.2.2), traitement dont n'ont pu bénéficier les autres lexiques.

Par-delà ces considérations, les écarts entre les résultats obtenus avec les différents lexiques montrent que les performances de FRMG dépendent de façon critique des informations fournies par le L*efff*. La nature du modèle utilisé, les TAG lexicalisées, rend cette conclusion peu surprenante, mais cette section a permis de la quantifier et de valider l'utilité du L*efff*. De plus, ces expériences nous ont permis d'améliorer la qualité du lexique verbal NewL*efff*, qui est actuellement sur le point de remplacer le lexique verbal actuellement intégré par défaut au L*efff*.

9.2 Informations morphologiques pour l'analyse syntaxique statistique en constituants : exploitation du L*effe* [8]

L'extraction de grammaires probabilistes à partir de corpus arborés est progressivement devenu le moyen le plus répandu pour fournir des règles de réécriture à un analyseur syntaxique. Toutefois, nous avons évoqué au début de ce chapitre (voir également la section A.10.2) la difficulté qu'il y a, notamment pour les langues à morphologie riche, à faire les bonnes généralisations à partir d'un ensemble d'exemples de taille réduite, et donc sujet à des problèmes de dispersion des données. L'étiquetage morphosyntaxique et la lemmatisation, deux traitements qui bénéficient grandement d'informations lexicales morphologiques, comme nous l'avons vu pour le premier au chapitre 8, sont l'une des approches permettant de diminuer l'impact de ce problème de dispersion des données. Nous avons donc procédé à plusieurs expériences visant à évaluer l'impact de tels traitements sur les performances d'analyseurs syntaxiques statistiques, à l'image de

8. Ce travail a été réalisé en collaboration avec Joseph Le Roux (Université Paris Nord) et Djamé Seddah (ALPAGE). Il est publié dans (Le Roux *et al.*, 2012b), et a de nombreux points communs avec nos travaux sur l'italien réalisé avec ces mêmes chercheurs (Seddah *et al.*, 2011, 2013a).

travaux tels que ceux de Goldberg et Tsarfaty (2008) et Goldberg et Elhadad (2013) sur l'hébreu. Nous avons conduit de telles expériences sur l'italien (Seddah *et al.*, 2011, 2013a) et l'espagnol (Le Roux *et al.*, 2012b). Le Le*fff* a été également utilisé à cette fin pour le français (Candito *et al.*, 2010 ; Seddah *et al.*, 2010a ; Green *et al.*, 2013) [9]. Nous nous concentrerons toutefois dans cette section sur nos expériences concernant l'analyse syntaxique en constituants de l'espagnol au moyen d'un analyseur non lexicalisé entraîné sur un corpus arboré de taille limitée.

Plus précisément, nos expériences reposent sur le lexique Alexina de l'espagnol, le Le*ff*e (cf. section 3.1.2 ; Molinero *et al.*, 2009b), qui contient environ 800,000 couples *(forme, catégorie*) distincts, et sur le corpus Cast3LB de l'espagnol, corpus en constituants de taille modeste, puisqu'il contient 3 500 arbres (cf. ci-dessous). L'espagnol est une langue à morphologie relativement riche, notamment dans son système verbal, rendant ainsi l'étiquetage morphosyntaxique complet relativement difficile. Cowan et Collins (2005) et Chrupała (2008) ayant déjà obtenu des résultats intéressants sur ce corpus, ils nous fournissent des points de comparaison, et notamment quant à la place des informations lexicales : en effet, ces deux travaux reposent sur un modèle d'analyse syntaxique lexicalisé, contrairement au travail décrit ici, qui repose sur un modèle non lexicalisé.

Le formalisme sur lequel repose l'analyseur syntaxique que nous avons utilisé, LORG (cf. plus bas), est en effet celui des grammaires non contextuelles probabilistes à annotations latentes (*Probabilistic Context-Free Grammars with Latent Annotations*, PCFG-LA ; Matsuzaki *et al.*, 2005 ; Petrov *et al.*, 2006). Ces grammaires diffèrent des grammaires non contextuelles probabilistes classiques en ceci que les symboles sont automatiquement affinés au cours de la phase d'entraînement, au moyen de techniques non supervisées. Elles ont été utilisées avec succès pour l'analyse syntaxique probabiliste d'un large éventail de langues, telles que le français (Candito et Seddah, 2010), l'allemand (Petrov et Klein, 2008), le chinois ou l'italien (Lavelli et Corazza, 2009).

Pour l'espagnol, les résultats décrits dans cette section constituaient l'état de l'art pour l'analyse en constituants, au moment de la publication de nos travaux, grâce notamment à l'étiqueteur MElt et au lexique Le*ff*e. Des expériences comparables sur l'italien (Seddah *et al.*, 2011, 2013a) nous ont également permis de dépasser l'état de l'art pour l'analyse en constituants de cette langue.

9.2.1 Corpus utilisé

Le corpus arboré Cast3LB (Castillian 3LB) (Civit et Martì, 2004) contient 3 509 phrases annotées en constituants et munies d'annotations fonctionnelles. Il est divisé comme

9. Ces expériences utilisent l'étiqueteur morphosyntaxique et lemmatiseur Morfette (Chrupała *et al.*, 2008), qui utilise le Le*fff* comme source d'informations lexicales.

habituellement en un sous-corpus d'entraînement (2 806 arbres couvrant 76 931 mots), un sous-corpus de développement (365 arbres) et un sous-corpus de test (338 arbres) [10].

Le jeu d'étiquettes du Cast3LB est riche. L'inventaire d'étiquettes morphosyntaxiques (parties du discours), en particulier, est très détaillé et encode des informations morphologiques précises. Les étiquettes non terminales, quant à elles, contiennent des informations sur la sous-catégorisation et les fonctions syntaxiques. Si l'on ignore le volet fonctionnel de ces étiquettes, on obtient 43 étiquettes non terminales distinctes. Le jeu d'étiquettes complet contient 149 symboles distincts.

9.2.2 Protocole expérimental

Nous avons réalisé différentes expériences à partir du Cast3LB afin d'évaluer l'impact de deux types de modifications sur la qualité de l'analyse syntaxique statistique telle que réalisée par un analyseur syntaxique PCFG-LA. La première, dont nous ferons état ici plus en détails, consiste à modifier des symboles pré-terminaux au moyen de jeux d'étiquettes morphosyntaxiques simplifiés. La seconde, pour laquelle nous renvoyons à (Le Roux *et al.*, 2012b), consiste à modifier les symboles terminaux en remplaçant les mots par leurs lemmes.

Nos expériences couvrent trois types d'entrées données à l'analyseur syntaxique :

Mots : texte sans aucune information morphosyntaxique,

Étiquettes de référence : texte avec étiquettes morphosyntaxiques issues du corpus arboré,

Étiquettes prédites : texte avec étiquettes morphosyntaxiques prédites par l'étiqueteur MElt présenté au chapitre précédent, entraîné sur le sous-corpus d'entraînement du Cast3LB et sur le lexique L*eff*e, et ce avec différents jeux d'étiquettes de granularités différentes [11].

Les différents jeux d'étiquettes que nous avons comparés sont les suivants :

— le jeu d'étiquettes *baseline*, qui est identique au jeu d'étiquettes utilisé par Cowan et Collins (2005) et Chrupała (2008) ; avec ce jeu d'étiquettes, le corpus d'entraînement contient 106 étiquettes distinctes ;

— le jeu d'étiquettes *reduced2*, qui est une simplification du jeu d'étiquettes *baseline* : nous ne gardons que les deux premiers caractères de chaque étiquette, ce qui élimine

10. Afin de pouvoir nous comparer directement avec les résulats de Chrupała (2008), nous avons appliqué à ce corpus les mêmes transformations que lui : des nœuds CP et SBAR sont ajoutées à la racine des propositions subordonnées et relatives. Aucune autre transformation n'a été appliquée, et notamment pas les modifications concernant les structures coordonnées utilisées par Cowan et Collins (2005).

11. Par rapport aux expériences du chapitre précédent, nous avons comparé l'utilisation par `megam` de l'algorithme de maximisation d'entropie et du perceptron multi-classes. Nous avons constaté que, sur les données dont il est question ici, les performances sont très similaires. Toutefois, comme évoqué au chapitre précédent, l'utilisation de l'algorithme du perceptron multi-classes permet un apprentissage bien plus rapide des modèles d'étiquetage, et ce sont ces modèles que nous avons utilisés pour ce travail.

Jeu d'étiquettes	***baseline***	***reduced2***	***reduced3***
Nombre d'étiquettes	106	42	57
Précision globale (%)	96,34	97,42	97,25
Précision sur les mots inconnus (%)	91,17	93,35	92,30

Tableau 9.4 – Résultats de MElt entraîné sur le sous-corpus d'entraînement du Cast3LB tels qu'évalués sur le sous-corpus de développement, pour chacun des trois jeux d'étiquettes. Le taux de mots inconnus dans le corpus de développement est de 13.5%.

des informations morphologiques mais conserve des catégories et sous-catégories majeures ; avec ce jeu d'étiquettes, le corpus d'entraînement contient 42 étiquettes distinctes ;

— le jeu d'étiquettes *reduced3*, qui est une variante de *reduced2* : contrairement à ce dernier, nous avons laissé dans *reduced3* les informations concernant le mode des verbes, ces informations ayant été identifiées par Cowan et Collins (2005) comme importantes pour l'analyse syntaxique ; avec ce jeu d'étiquettes, le corpus d'entraînement contient 57 étiquettes distinctes.

Les résultats en termes de précision des différents modèles MElt ainsi obtenus, tels que mesurés sur le sous-corpus de développement du Cast3LB, sont fournis à la table 9.4. Les scores de précision globaux y sont accompagnés des scores de précision sur les seuls mots inconnus, c'est-à-dire les mots absents du sous-corpus d'entraînement, soit 13,5% des mots du sous-corpus de développement.

9.2.3 L'analyseur syntaxique LORG

L'analyseur syntaxique que nous avons utilisé est LORG (Attia *et al.*, 2010) [12], analyseur PCFG-LA qui repose sur l'algorithme CKY. Toutes nos expériences ont été reproduites avec deux types de modèles différents pour la gestion des mots inconnus :

— des modèles où les mots inconnus sont remplacés par un marqueur unique de mot inconnu, *UNK* (*modèles génériques*) ;

— des modèles les mots inconnus sont typés grâce aux suffixes pertinents identifiés préalablement et grâce à leurs propriétés typographiques : les suffixes qui aident à discriminer les parties du discours sont extraits du corpus d'apprentissage et ordonnés selon l'importance de leur contribution à la tâche de prédiction des étiquettes morphosyntaxiques Attia *et al.* (2010).

12. Distribué à l'adresse suivante : `https://github.com/CNGLdlab/LORG-Release`.

Dans cette section, nous ne ferons mention que des résultats de modèles génériques. Le lecteur intéressé pourra se référer à (Le Roux *et al.*, 2012b) pour des résultats complets, y compris avec le second type de modèles.

Les grammaires ont été produites au moyen de l'algorithme de Petrov et Klein (2007) avec 3 itérations de la séquence split/merge/smooth [13]. Concernant les règles lexicales, nous avons appliqué la stratégie nommée *simple lexicon* dans l'analyseur de Berkeley. Les mots rares (moins de 3 occurrences dans les données d'apprentissage) sont traités de la même façon que le sont les mots inconnus dans les textes à analyser.

9.2.4 Résultats

Les premières expériences que nous avons menées l'ont été avec le jeu d'étiquettes *baseline*. Les résultats en sont résumés dans la partie supérieure de la table 9.5, qui présente les scores obtenus sur le sous-corpus de développement du Cast3LB dans les trois configurations mentionnées ci-dessus : texte brut, catégories de référence et catégories prédites [14]. Ces scores sont calculés en ignorant la ponctuation, comme il est d'usage, et en ne prenant en compte que les phrases de 40 mots au plus, afin de rester comparables aux travaux antérieurs [15].

Unités analysées	Précision	Rappel	F-mesure	Exact	Étiqu. morphosynt.
Jeu d'étiquettes *baseline*					
Mots	81,42	81,04	**81,23**	**14,47**	**90,89**
Étiquettes de référence	87,83	87,49	**87,66**	**30,59**	99,98
Étiquettes prédites	84,47	84,39	**84,43**	**22,44**	95,82
Jeu d'étiquettes *reduced2*					
Mots	78,86	79,02	78,94	15,23	88,18
Étiquettes de référence	86,56	85,90	86,23	26,64	100,00
Étiquettes prédites	84,16	83,81	83,99	21,05	**96,76**
Jeu d'étiquettes *reduced3*					
Mots	79,61	79,78	79,69	**14,90**	87,29
Étiquettes de référence	88,08	87,69	**87,89**	**30,59**	100,00
Étiquettes prédites	85,56	85,38	**85,47**	23,03	96,56

Tableau 9.5 – Résultats d'analyse syntaxique sur le sous-corpus de développement du Cast3LB avec les trois jeux d'étiquettes ($\leq$ 40 mots). « Exact » indique le pourcentage de phrases dont l'analyse est intégralement correcte.

13. Nous avons également essayé d'appliquer 4 et 5 itérations, mais il s'est avéré que les meilleurs résultats, sur ce corpus, étaient obtenus avec 3 itérations.

14. Si l'analyseur ne peut construire une analyse syntaxique respectant l'étiquetage morphosyntaxique donné en entrée (lorsque c'est le cas), il ignore cet étiquetage (il procède alors lui-même à l'étiquetage, comme dans le cas *Mots*)

15. Tous nous les scores prennent en compte l'étiquette des constituants ; ils ont été obtenus au moyen de l'outil `Parseval`.

Comme indiqué précédemment, ce jeu d'étiquettes contient 106 étiquettes distinctes. Cela signifie d'un côté que ces étiquettes véhiculent des informations précises, mais d'un autre côté la dispersion des données est élevée.

Les résultats obtenus avec les jeux d'étiquettes *reduced2* et *reduced3* sont indiqués dans les deux dernières sections du tableau 9.5. On constate que les résultats sont meilleurs avec le jeu d'étiquettes *reduced3*, ce qui montre que le mode verbal est une information importante pour l'analyse des verbes au niveau syntaxique, mais que d'un autre côté un jeu d'étiquettes trop fin dégrade le résultat final, vraisemblablement parce qu'il fait prendre au modèle plus simple d'étiquetage morphosyntaxique des décisions qu'il aurait mieux fallu laisser à l'analyseur syntaxique.

9.2.5 Discussion

Peu de travaux ont été publiés sur l'analyse syntaxique statistique de l'espagnol. Le travail initial de Cowan et Collins (2005) consistait en une étude approfondie de l'impact de différents traits morphologiques sur l'analyse syntaxique lexicalisée au moyen du modèle 1 de Collins et sur le gain obtenu grâce au réordonnanceur de Collins et Koo (2005) utilisé avec un jeu de traits développé pour l'anglais. Une comparaison directe avec leurs résultats est difficile, car leur corpus d'évaluation n'est pas le même que le nôtre. Ils obtiennent une f-mesure de 85,1% sur les phrases de 40 mots ou moins.

Nous sommes toutefois directement comparables avec Chrupała (2008), qui a adapté le modèle 2 de Collins à l'espagnol, à condition d'ignorer comme lui, lors de l'évaluation, les nœuds CP et SBAR rajoutés initialement [16]. En utilisant la même segmentation du corpus en sous-corpus d'entraînement, de développement et de test, si l'on fournit en entrée à l'analyseur syntaxique l'étiquetage morphosyntaxique de référence, notre analyseur atteint une f-mesure supérieure à celui de Chrupała (2008) d'environ 2,3% absolus (cf. table 9.6 pour les résultats sur le sous-corpus de test [17]).

Lorsque le corpus d'entraînement est très petit, comme c'est le cas ici, les PCFG-LA manquent cruellement de données annotées. Ce n'est qu'en réduisant significativement la taille du jeu d'étiquettes et en utilisant un étiqueteur morphosyntaxique performant (et/ou un lemmatiseur, mais nous ne l'avons pas décrit ici) que l'on peut atteindre de bons résultats, meilleurs que ceux publiés précédemment. La sensibilité des PCFG-LA au problème de la dispersion des données, et notamment des données lexicales, a été également mise en avant et étudiée, sur le français, par Seddah *et al.* (2009). Ils ont montré que les performances des analyseurs lexicalisés en constituants de niveau état-

16. Cela dit, parce qu'il s'intéressait à l'induction de grammaires LFG à large couverture, Chrupała (2008) a enrichi le schéma d'annotation pour rajouter des chemins fonctionnels sur les non-terminaux, plutôt que de chercher à obtenir le jeu d'étiquettes optimal en termes d'analyse syntaxique proprement dite.

17. Nous avons obtenus des résultats légèrement meilleurs avec le modèle de gestion des mots inconnus reposant sur les suffixes pertinents, comme évoqué précédemment. On pourra se reporter à (Le Roux *et al.*, 2012b) pour plus de détails.

Jeu d'étiquettes	Mode	Terminaux	F-mesure selon les phrases évaluées		
			Toutes	$\leq$ 70 mots	$\leq$ 40 mots
reduced3	Generic	Étiquettes prédites	83,92	84,27	85,08
évaluation sans CP et SBAR			*84,02*	*84,37*	*85,24*
reduced3	Generic	Étiquettes de référence	86,21	86,63	87,84
évaluation sans CP et SBAR			*86,35*	*86,77*	*88,01*
baseline	(Chrupała, 2008)	Étiquettes de référence			
évaluation sans CP et SBAR			*83,96*	*84,58*	–

Tableau 9.6 – Résultats (f-mesures) sur le sous-corpus de test du Cast3LB

de-l'art (modèles de Charniak, Collins, etc.) croisent celles des analyseurs non lexicalisés de type Berkeley (comme LORG) lorsque le corpus d'entraînement contient entre 2 500 et 3 000 phrases. Ici, avec environ 2 800 phrases d'entraînement, nous sommes donc vraisemblablement dans une zone où les deux types d'analyseurs obtiennent des résultats comparables, l'espagnol et le français étant relativement proches par leurs propriétés syntaxiques.

Il est donc satisfaisant d'obtenir des performances de niveau état-de-l'art simplement grâce à l'exploitation d'outils d'étiquetage morphosyntaxique et de lemmatisation, outils qui reposent eux-mêmes sur des informations lexicales morphologiques. Cela démontre que les informations morphologiques que peut apporter un lexique externe à un analyseur statistique en constituants tel que LORG, soit directement soit au travers d'un étiqueteur morphosyntaxique, sont de nature à améliorer la qualité du modèle probabiliste et donc celle des analyses produites. Comme nous allons le voir à la section suivante, il en est de même pour les informations syntaxiques, bien que l'intégration de telles informations soit plus délicate à mettre en œuvre.

9.3 Informations syntaxiques pour l'analyse syntaxique statistique en dépendances : comparaison du Le*fff* avec d'autres types d'informations lexico-syntaxiques [18]

Comme nous l'avons discuté au début de ce chapitre, l'intégration d'informations lexicales de niveau syntaxique à un analyseur statistique est une problématique à la fois importante, difficile, et encore ouverte. Nous nous sommes penchés sur cette problématique pour le cas du français, en nous restreignant à la sous-catégorisation verbale. Cette étude a ainsi pour objectif d'apporter des éléments de réponse aux deux

18. Le travail présenté dans cette section a été réalisé en collaboration avec Seyed Abolghasem Mirroshandel (alors ancien post-doc au LIF, Université de Marseille et chercheur à l'Université de Guilan, en Iran) et Alexis Nasr (professeur à l'Université de Marseille). Il est publié dans (Mirroshandel *et al.*, 2013).

questions suivantes : (i) comment intégrer des informations lexico-syntaxiques à un tel analyseur, et (ii) quelles sources d'informations lexico-syntaxiques s'avèrent les plus utiles pour améliorer les performances de l'analyseur ?

Concernant le premier point, l'approche la plus courante est l'utilisation d'un réordonnanceur qui intègre ce type de contraintes, comme nous l'avons rappelé en introduction de ce chapitre (Collins, 2000 ; Charniak et Johnson, 2005 ; Versley et Rehbein, 2009). Il s'agit d'une architecture en deux étapes : (i) tout d'abord, on demande au modèle probabiliste de base de fournir ce qu'il identifie comme les n meilleures analyses [19] ; ensuite, (ii) un réordonnanceur réévalue la pertinence de chacun des n arbres en fonction des informations lexico-syntaxiques disponibles, processus qu'il aurait été trop coûteux de réaliser pendant la première phase sur l'ensemble des arbres possibles ou, *a fortiori*, au cours du processus d'analyse lui-même. L'un des problèmes avec cette approche est qu'il est possible qu'aucun des n meilleurs arbres n'attribuent le bon cadre de sous-catégorisation à tous les verbes de la phrase. Nous avons donc étudié la possibilité de recombiner des sous-parties de plusieurs de ces n arbres, c'est-à-dire des sous-analyses, pour construire une analyse qui peut ne pas en faire partie mais qui est maximalement cohérente avec les contraintes de sous-catégorisation.

Concernant la deuxième des questions mentionnées ci-dessus, nous avons étudié l'impact de l'intégration d'informations lexico-syntaxiques provenant de trois types de sources : le lexique Le*fff*, le corpus d'entraînement (FTB-TRAIN) et un grand corpus analysé automatiquement par la version de base de l'analyseur.

Nous passerons successivement en revue les caractéristiques de l'analyseur utilisé, la façon dont nous avons extrait et représenté les informations lexico-syntaxiques à partir des trois types de sources mentionnés ci-dessus, puis deux différentes façons d'exploiter ces informations dans l'analyse syntaxique en suivant l'idée générale de la recombinaison de sous-analyses.

9.3.1 L'analyseur syntaxique MATE

L'analyseur syntaxique utilisé dans ces expériences est MATE, réimplémentation et amélioration par Bohnet (2010) de l'analyseur syntaxique en dépendances de McDonald *et al.* (2005), qui s'appuie sur une représentation en graphes des structures de dépendances possibles [20]. Nous avons entraîné cet analyseur sur le FTB, en suivant la même répartition

19. Ces n meilleures analyses peuvent être fournies sous forme de liste d'arbres ou, de façon plus compacte, sous forme de forêt partagée. Dans ce dernier cas, la méthode la plus simple est de partir de la forêt complète et d'en retirer les branches ne participant à aucune des n meilleures analyses. Toutefois, la forêt résultant de cet élagage peut contenir bien plus de n arbres. Nous avons étudié ce problème dans le cas de l'analyse en constituants avec une PCFG et avons proposé un algorithme permettant de construire une forêt ne contenant que les n meilleurs arbres, au prix d'une défactorisation partielle de la forêt (Boullier *et al.*, 2009).

20. Les motivations principales de ce choix sont d'une part les bonnes performances de cet analyseur et d'autre part le fait qu'il soit possible de lui appliquer des contraintes structurelles sur les analyses

en sous-corpus d'entraînement, de développement et de test que dans les expériences précédentes effectuées sur ce corpus arboré.

L'analyseur obtenu conduit à des résultats proches de l'état de l'art de l'analyse automatique du français, comme évoqué à la section 9.1.1 et montré plus en détail à la table 9.7 en termes de score d'attachement étiqueté (LAS) et de score d'attachement non étiqueté (UAS). Nous avons également défini une mesure spécifique à ce travail, le score de précision valencielle (*Sub-categorization Accuracy Score*, SAS), défini comme étant la proportion d'occurrences verbales auxquelles l'analyse retenue attribue le cadre de sous-catégorisation correct.

	FTB-DEV	FTB-TEST
SAS	79,88	80,84
LAS	88,53	88,88
UAS	90,37	90,71

TABLEAU 9.7 – Résultats de MATE sur le FTB (étiquettes morphosyntaxiques de référence fournies en entrée). On se reportera au texte pour les définitions des métriques employées.

9.3.2 Informations lexico-syntaxiques

Comme indiqué précédemment, nous avons utilisé et comparé trois sources d'informations lexico-syntaxiques : le Le*fff*, le corpus d'entraînement et un grand corpus analysé automatiquement par l'analyseur de base dont nous venons de décrire les performances. Dans ce travail, les informations lexico-syntaxiques sont encodées sous la forme de cadres de sous-catégorisation. Ces derniers, contrairement à ce qui se passe dans Alexina, ne contiennent aucune factorisation ou alternative : il s'agit de cadres totalement instanciés, y compris par le lemme verbal et l'étiquette morphosyntaxique du verbe [21]. Pour cette raison, et pour les distinguer des cadres à la Alexina, nous parlerons de cadres de sous-catégorisation ancrés, ou CSCA. Les arguments eux-mêmes sont définis par une fonction syntaxique du FTB (cf. section 9.1.2) et l'étiquette morphosyntaxique de la tête de l'argument (en passant par dessus les prépositions pour les arguments prépositionnels, et en fusionnant réalisations nominales et pronominales en une même étiquette `N`). La figure 9.8 donne trois CSCA pour le lemme *donner*, avec à chaque fois un exemple instanciant le CSCA.

fournies (Mirroshandel et Nasr, 2011), fonctionnalité qui nous permettra de demander à l'analyseur de respecter certaines contraintes liées aux sous-catégorisations verbales.

21. Rappelons que dans le jeu d'étiquettes utilisé, le jeu d'étiquettes utilisé par MElt_{fr}, un verbe peut être étiqueté VINF (infinitif), VPP (participe passé), VPR (participe présent) ou V (forme finie).

CSCA	Exemple
(donner,(V,(suj,N),(obj,N)))	*Jean donne un livre*
(donner,(V,(suj,N),(obj,N),(a_obj,N)))	*Jean donne un livre à Marie*
(donner,(VPP,(suj,N),(aux_pass,V), (a_obj,N),(p_obj,N)))	*Le livre est donné à Marie par Jean*

Tableau 9.8 – Trois CSCA pour le verbe *donner*

9.3.2.1 Extraction des trois ensembles de CSCA

Les CSCA n'étant pas directement modélisés par le modèle d'analyse syntaxique extrait du corpus d'entraînement, FTB-train, en raison du caractère relativement local des informations qu'il contient, il fait sens d'en extraire explicitement les CSCA qui y sont attestés. Ce processus est direct : on extrait de chaque arbre les annotations fonctionnelles associées à chaque occurrence verbale. Les résultats de cette extraction sont fournis dans la colonne FTB-train de la table 9.9, ce lexique de CSCA étant nommée T par la suite [22].

Afin d'obtenir une meilleure couverture qu'avec le seul FTB-train, et au risque de diminuer la qualité, nous avons également extrait un ensemble de CSCA à partir du résultat de l'analyse syntaxique automatique d'un corpus brut d'environ 2 millions de mots couvrant des genres divers [23, 24]. Nous avons utilisé MElt_{fr} pour l'étiquetage en parties du discours, MACAON (Nasr *et al.*, 2011) pour la lemmatisation et la version de base de l'analyseur utilisé ici pour l'analyse syntaxique. L'extraction des CSCA à partir des analyses est alors immédiate, le format obtenu étant identique au corpus d'entraînement FTB-train. Nous ne conservons alors que les CSCA apparaissant au moins 10 fois (cf. tableau 9.9, colonne A_{10}). Notons que deux sources d'erreurs apparaissent dans les CSCA ainsi extraits : les erreurs d'étiquetage ou de lemmatisation, qui conduisent à des CSCA mal ancrés, et les erreurs d'analyse, qui conduisent à des CSCA structurellement incorrects.

Enfin, nous avons converti le lexique verbal extensionnel du Le*fff* [25] en un lexique de CSCA [26]. Le résultat, à propos duquel des informations quantitatives sont fournies à la

22. Il peuvent être comparés à la ressource TreeLex (Kupść, 2008), évoquée au chapitre 5. La différence entre ces deux ressources réside dans le fait que Treelex utilise une variante plus abstraite de la notion de cadre de sous-catégorisation, d'où seulement 58 cadres verbaux contre 666 dans notre cas, et en conséquence moins de cadres par lemme verbal (1,72 contre 4,83).

23. Plus précisément, nous avons rassemblé 2 millions de phrases issues de dépêches AFP, 3 millions de phrases du corpus de l'Est Républicain et 1,6 millions de phrases de la wikipedia française, pour un total de 150 millions de mots.

24. Il s'agit donc d'une procédure d'acquisition automatique de lexique syntaxique à partir de corpus brut et au moyen d'une analyse syntaxique automatique, au sens de Briscoe et Carroll (1997), comme discuté à la section chapitre A.6. Cependant, l'accent n'est pas mis ici sur cette procédure en tant que telle, mais sur l'intégration des informations ainsi extraites dans l'analyseur.

25. La version utilisée contenait 10 618 entrées verbales intensionnelles pour 7 835 lemmes verbaux distincts.

26. Pour chaque lemme verbal, nous avons ainsi extrait les cadres de sous-catégorisation associés à un représentant de chacune des quatre étiquettes morphosyntaxiques possibles (`V`, `VINF`, `VPR` et `VPP`), cadres que

Source	FTB-train	Le*fff*	Corpus analysé automatiquement
Seuillage	aucun	aucun	occ. ≥ 10
Nom de la ressource	T	L	A_{10}
Lemmes verbaux distincts	2 058	7 824	3 923
CSCA distincts	666	1 469	1 355
CSCA par verbe (moyenne)	4,83	52,09	13,45

Tableau 9.9 – Données quantitatives sur les ressources lexico-syntaxiques extraites

colonne Le*fff* de la table 9.9, est une ressource que nous nommerons L par la suite et qui contient pas moins de 810 246 CSCA.

9.3.2.2 Couverture des ensembles de CSCA extraits

Nous avons évalué la couverture des différents lexiques de CSCA sur le FTB-dev au moyen de deux mesures : la couverture lexicale (pourcentage de verbes présents dans le lexique de CSCA par rapport à ceux attestés dans le FTB-dev) et la couverture syntaxique (pourcentage de CSCA présents dans le lexique par rapport à ceux attestés dans le FTB-dev). Les résultats sont fournis à la table 9.10, calculés à la fois pour les occurrences et sur les types. Ils montrent que la couverture lexicale de nos lexiques de CSCA est bonne, puisqu'elle va sur les types de 89,56 (ressource T) à 98,08 (ressource A_{10}) et sur les occurrences de 96,98 (T) à 99,85 (L). Que T ait la moins bonne couverture lexicale n'est pas surprenant puisqu'il s'agit d'un lexique extrait d'un corpus de taille modeste [27]. La couverture syntaxique, par définition inférieure à la couverture lexicale, est maximale pour A_{10} [28]. Les scores obtenus par L sont eux-mêmes bien plus élevés que ceux de T.

La couverture de L, lexique de CSCA issu du Le*fff*, est basse par rapport à ce à quoi on aurait pu s'attendre. Nous avons identifié au moins quatre facteurs ayant conduit à cela : (i) des erreurs dans le FTB-dev, (ii) des erreurs dans le processus d'extraction des CSCA à partir du FTB-dev, conduisant à des CSCA erronnés dont il est normal qu'ils ne soient pas présents dans L, (iii) des erreurs dans le processus de conversion du Le*fff* en le lexique L de CSCA, et (iv) des erreurs dans le Le*fff*. Afin de mieux comprendre ce qui se passe, nous avons analysé manuellement un petit extrait de 30 CSCA extraits du FTB-dev mais non couverts par L, choisis aléatoirement parmi les 513 CSCA qui sont dans ce cas. Il se trouve

nous avons ensuite défactorisés (un cadre du Le*fff* correspondant parfois à des dizaines de CSCA) et dont nous avons converti les jeux de fonctions syntaxiques et de réalisations, afin de respecter la définition des CSCA.

27. Par ailleurs, des expériences faites sur les CSCA extraites du corpus annotées automatiquement avec d'autres seuls d'occurrence que 10 (A_{10}) montrent que la couvrture ne diminue que peu lorsque i augmente (Mirroshandel et Nasr, 2011) : en absence de tout filtre, les CSCA extraites étant relativement bruitées, le filtrage par nombres d'occurrences permet de réduire la taille du lexique sans quasiment diminuer sa couverture, et donc de réduire le bruit.

28. Le seuillage sur les occurrences a ici un impact plus grand que sur la couverture lexicale. Cf. (Mirroshandel et Nasr, 2011).

que seuls 6 de ces 30 CSCA sont corrects et sont effectivement la trace de manques dans le Le*fff* ou d'erreurs dans le processus de construction de L à partir du Le*fff*. Les autres CSCA absents de L sont tous erronés [29].

		T	L	A_{10}
Couverture lexicale	types	89,56	99,52	98,08
	occ.	96,98	99,85	99,50
Couverture syntaxique	types	62,24	78,15	88,84
	occ.	73,54	80,35	92,39

Tableau 9.10 – Couvertures lexicale et syntaxique des lexiques de CSCA sur le FTB-dev

9.3.3 Prise en compte des lexiques de CSCA dans l'analyseur syntaxique

Comme indiqué précédemment, l'architecture que nous avons mise en place consiste à appliquer dans un premier temps l'analyseur syntaxique probabiliste de base en lui demandant de ne conserver que les n meilleurs arbres, puis d'exploiter un lexique de CSCA pour construire une nouvelle analyse à partir de sous-analyses extraites de ces n meilleurs arbres de base. Nous faisons l'hypothèse, qui est également une limitation de ce travail, que le CSCA correct pour chaque occurrence verbale dans une phrase est attesté dans au moins l'une des n meilleures analyses de base pour cette phrase. Nous avons estimé la marge de progression maximale en SAS atteignable par une telle approche, en faisant usage d'un oracle sur le FTB-dev, pour $n = 100$. Le résultat est un SAS maximal atteignable de 95,16%, à comparer aux 79,88% de l'analyseur de base sur le FTB-dev.

Nous avons étudié deux façons d'intégrer un lexique de CSCA dans l'analyseur syntaxique utilisé ici. La première méthode, dite *Post-Processing* (PP) consiste à choisir pour chaque occurrence verbale son CSC le plus vraisemblable à partir de l'étude du lexique et des 100 meilleures analyses puis à modifier le cas échéant le meilleur arbre de telle sorte qu'il respecte les CSC ainsi retenus. Mais le résultat peut être incohérent, ces modifications locales pouvant, dans leur voisinage, induire des configurations bien peu vraisemblables. La seconde méthode, dite *Double Parsing* (DP) s'appuie donc sur les résultats de la première pour demander à l'analyseur de produire la meilleure analyse possible parmi celles qui respectent l'ensemble des modifications proposées par PP, utilisant pour cela la technique proposée par Mirroshandel et Nasr (2011). Nous ne

29. Parmi eux, nombreux sont ceux qui ont des propriétés invalides faciles à identifier automatiquement : parmi 513 CSCA extraits du FTB-dev mais absents de L, 33 ont 2 sujets, 59 ont un objet direct dont la tête est une préposition et 84 n'ont pas de sujet bien qu'ils correspondent à des formes finies. De telles erreurs sont soit des erreurs du FTB-dev soit des erreurs du processus d'extraction de CSCA à partir du FTB-dev. On notera que de telles erreurs du processus d'extraction à partir du FTB-dev ont également créé des CSCA incorrects à partir du FTB-train ou du corpus analysé automatiquement, améliorant ainsi de façon artificielle la couverture syntaxique de T et de A_{10}.

décrirons pas ici ces deux techniques en détail. On pourra se référer à cette fin à (Mirroshandel *et al.*, 2013).

Les résultats obtenus sur FTB-TEST sont donnés à la table 9.11. On constate que PP permet déjà une amélioration significative du SAS : PP corrige certaines erreurs de sous-catégorisation faites par l'analyseur de base. Mais cette amélioration est bien moindre que le maximum théorique atteignable, tel que mesuré précédemment avec un oracle. De plus, les scores LAS et UAS ne sont que très faiblement améliorés. Ceci est lié au fait que le nombre de dépendances modifiées par PP est très faible relativement à l'ensemble des dépendances. Les meilleurs résultats sont obtenus avec T. Ceci est probablement lié au fait que cette ressource soit très proche des données d'évaluation, et ce malgré sa couverture plus basse que les autres lexiques de CSCA. Les SAS obtenus avec DP sont les mêmes qu'avec PP, puisque les CSCA sont inchangés. Comme attendu, les scores LAS et UAS sont en revanche meilleurs avec DP qu'avec PP. Recalculer une solution optimale globalement qui respecte les décisions concernant les CSCA a ainsi amélioré la précision sur dépendances « voisines » des occurrences verbales : il y a un effet boule de neige, la correction des dépendances verbales conduisant désormais à des corrections sur les autres dépendances. On constate également que DP conduit à des résultats de LAS et de UAS très similaires pour les trois lexiques de CSCA.

MESURE	ANALYSEUR DE BASE	PP(T)	PP(L)	PP(A_{10})	DP(T)	DP(L)	DP(A_{10})
SAS	80,84	**83,11**	82,14	82,17	**83,11**	82,14	82,17
LAS	88,88	**89,14**	89,03	89,03	**89,30**	**89,25**	**89,31**
UAS	90,71	**90,91**	**90,81**	**90,82**	**91,07**	**91,05**	**91,08**

TABLEAU 9.11 – Résultats sur FTB-TEST de l'application de PP et de DP comparés aux résultats de l'analyseur de base

9.3.4 Discussion

Bien que cette expérience puisse être améliorée de plusieurs façons [30], nous avons montré qu'il était possible d'améliorer les performances d'un analyseur syntaxique statistique en dépendances au moyen d'informations lexico-syntaxiques, y compris d'informations extraites d'un lexique comme le Le*fff*. De plus, cette amélioration est localisée en partie sur les cadres de sous-catégorisation, qui sont des informations

30. Les améliorations que l'on peut apporter à une telle étude sont nombreuses. Le processus d'extraction des lexiques de CSCA est améliorable, notamment pour l'extraction de L à partir du Le*fff*. En effet, ce processus inclut une véritable opération de conversion, les informations lexico-syntaxiques devant être mises en cohérence avec les choix d'annotation du FTB. Par ailleurs, des premiers résultats faisant usage de combinaisons des différentes ressources lexico-syntaxiques utilisées ici n'ont pas montré d'amélioration importante par rapport à l'utilisation des ressources individuelles (Mirroshandel *et al.*, 2013). Mais des approches plus fines pour la combinaison des lexiques de CSCA peuvent être envisagées, qui pourraient notamment passer par une pondération de L au moyen des comptes présents dans les autres ressources.

particulièrement pertinentes en vue d'une analyse plus profonde (par exemple, sémantique) et en même temps difficiles à modéliser convenablement avec les approches les plus utilisées pour l'analyse statistique en dépendances.

9.4 Décalage entre tokens et formes et analyse syntaxique [31]

Les difficultés que nous avons abordées au chapitre précédent à propos de l'étiquetage morphosyntaxique de textes bruités issus du web ne disparaissent pas, bien au contraire, lorsque l'on cherche à en faire l'analyse syntaxique automatique. Dans de telles données, non seulement les tokens altérés sont légions, mais les structures syntaxiques elles-mêmes ne correspondent pas nécessairement toujours à ce que l'on rencontre dans les textes édités que l'on trouve dans la plupart des corpus arborés, et notamment l'usage de l'impératif et du discours direct (Foster *et al.*, 2011a,b ; Gimpel *et al.*, 2011 ; Elsner et Charniak, 2011). Les difficultés principales avaient déjà été décrites notamment dans le travail séminal de Foster (2010).

Naturellement, ces difficultés ne sont pas les seules qui contribuent à rendre délicate la tâche d'analyse syntaxique de par la non-correspondance entre tokens et formes. Un autre problème, très étudié également, est naturellement celui des « unités multi-mots ». Nous n'avons pas travaillé spécifiquement sur ce sujet, mais on pourra se reporter notamment à aux travaux de Constant et Sigogne (2011), Constant *et al.* (2012), Green *et al.* (2013), Vincze *et al.* (2013), Constant *et al.* (2013), Le Roux *et al.* (2014) ou, sur les tweets, de Kong *et al.* (2014). Ce thème de recherche reste néanmoins ouvert, et on peut s'attendre à des progrès grâce à une meilleure exploitation de ressources lexicales construites manuellement voire (semi-)automatiquement.

Dans cette section, nous nous concentrerons néanmoins, comme à la section 8.4, sur la problématique de l'analyse des textes bruités, notamment ceux que l'on trouve sur le web, mais cette fois-ci au niveau de l'analyse syntaxique. Une stratégie telle que celle mise en place pour l'étiquetage morphosyntaxique reste aujourd'hui à concevoir et à développer. La raison en est que, contrairement aux étiquettes morphosyntaxiques qu'il est possible, comme nous l'avons vu, de reporter sur les tokens d'origine, il est très délicat de transformer des arbres dont les feuilles seraient les formes normalisées en des arbres ayant pour formes les tokens d'origine. En effet, les tokens amalgamés tels qu'en français `chépa` (pour *je (ne) sais pas*) ou `qil` (pour *qu'il*) ne correspondent pas à des sous-arbres de l'arbre syntaxique complet. Pour cette raison, nous nous sommes concentrés sur l'étude de l'impact de nos résultats en étiquetage morphosyntaxique sur l'analyse syntaxique

31. Le travail présenté dans cette section a été réalisé en collaboration. Le développement du French Social Media Bank a été réalisée principalement avec Djamé Seddah (Seddah *et al.*, 2012c,d). Notre participation à la campagne SANCL 2012, qui nous a valu d'être classés deuxièmes (Petrov et McDonald, 2012), a est également le résultat d'une collaboration avec Djamé Seddah, avec des contributions de Marie Candito et Joseph Le Roux.

(cf. section 9.4.1), éventuellement complétés par un clustering lexical pour diminuer la dispersion des données (cf. section 9.4.2). La section 9.4.1 décrit les résultats obtenus dans le cadre du développement du FSMB, alors que la section 9.4.2 détaille l'architecture que nous avons mise en place dans la cadre de la campagne SANCL d'évaluation des analyseurs syntaxiques sur des données de l'anglais issues du web.

9.4.1 Analyse syntaxique de textes bruités : expériences préliminaires sur le French Social Media Bank

Après avoir annoté le FSMB au niveau morphosyntaxique, comme discuté à la section 8.4.2, l'étape suivante du développement de ce corpus annoté était son annotation en constituants. Comme indiqué dans cette même section, nous avons à nouveau procédé par pré-annotation puis correction manuelle. Un exemple de phrase complètement annotée (étiquettes morphosyntaxiques, constituants, mais également annotations fonctionnelles) est montré à la figure 9.1.

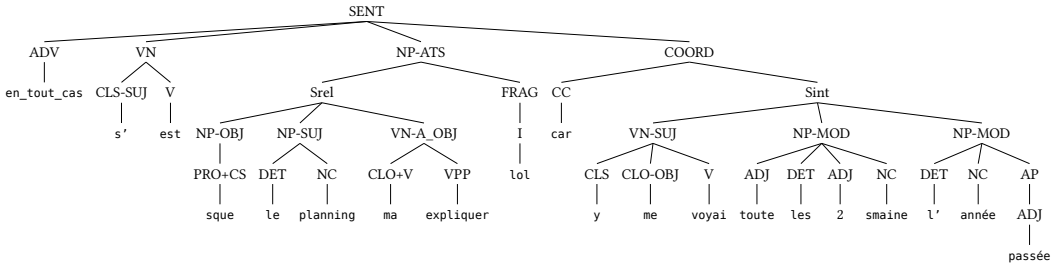


FIGURE 9.1 – Exemple issu du French Social Media Bank (sous-corpus Doctissimo/B+).

Comme au niveau morphosyntaxique, les conventions d'annotation des structures en constituants sont quasiment identiques à celles du FTB. Trois modifications importantes sont néanmoins à signaler à ce niveau. D'une part, nous avons homogénéisé l'annotation des constituants projetés par les prépositions et les complémenteurs [32]. D'autre part, nous avons utilisé un type de constituants supplémentaire, FRAG, qui permet d'identifier des constituants qui ne s'attachent pas syntaxiquement à la clause principale [33]. Enfin, nous avons pris en compte les amalgames non standard de la façon suivante : lorsque les formes sous-jacentes à un amalgame font partie du même constituant immédiat, nous utilisons la même structure en constituants que celle qui aurait été attribuée à l'équivalent édité. Si ce n'est pas le cas, nous faisons usage de structures localement plates.

32. Dans le FSMB, les prépositions projettent toutes un syntagme prépositionnel (PP), y compris celles introduisant une infinitive, contrairement aux conventions utilisées pour le FTB. Les complémenteurs, quant à eux, ont tous pour nœud frère à leur droite un nœud clause, à l'inverse de la structure plate utilisée par le FTB.

33. Les constituants FRAG recouvrent notamment les salutations, les *time stamps* et les smileys, mais aussi les *at-mentions*, noms d'utilisateurs et URL parfois disposés à la fin d'une phrase ou d'un *post*.

L'analyseur syntaxique que nous avons utilisé pour l'étape de pré-annotation en constituants est l'analyseur de Berkeley (Petrov *et al.*, 2006) adapté au français par Crabbé et Candito (2008), notamment pour une prise en compte raisonnable de la morphologie du français, et entraîné sur le FTB-UC [34]. Les étiquettes morphosyntaxiques précédemment obtenues sont données en entrée à l'analyseur. Dans les cas où l'analyseur ne peut faire autrement que d'ignorer certaines étiquettes morphosyntaxiques pour arriver à produire une analyse, les étiquettes de référence sont réinsérées *a posteriori*.

Après que les annotatrices ont corrigé les résultats de la pré-annotation, nous avons pu évaluer la qualité de ladite pré-annotation. Les résultats que nous avons obtenus sur les différents sous-corpus du FSMB sont fournis à la table 9.12, ainsi que ceux obtenus par le même analyseur sur le FTB-UC par Candito et Seddah (2010), à des fins de comparaison. Ils reposent sur la métrique PARSEVAL telle qu'implémentée dans la version de Petrov et McDonald (2012) de l'outil classique `evalb`, version que nous appellerons `evalb`$^+$ et que nous utiliserons également dans la section suivante. Ils évaluent la performance de l'analyseur sur les constituants étiquetés. Ces résultats pourraient être améliorés, notamment par l'emploi des techniques dont il sera fait usage à la section suivante sur l'anglais. Mais notre objectif était de fournir de premiers résultats en analyse syntaxique de données bruitées issues du web, ainsi que de diminuer le temps de travail manuel pour le développement du FSMB.

On pourra noter que les données issues du web qui constituent le FSMB semblent encore plus difficiles à analyser en constituants, du moins avec l'analyseur de Petrov *et al.* (2006) et le FTB-TRAIN comme corpus d'entraînement, que les données biomédicales du sous-corpus EMEA du corpus arboré Sequoia, données pourtant remarquables par leur nombre élevé d'inconnus et de structures inhabituelles (Candito *et al.*, 2011). Les scores en étiquetage morphosyntaxique, qui évaluent l'étiquetage choisi par l'analyseur, sont souvent significativement moins bons que les scores obtenus automatiquement à la section 8.4.2, et ce malgré le fait que les étiquettes fournies à l'analyseur ont toutes été corrigées manuellement. Cela illustre la difficulté de la tâche d'analyse de telles données. Le sous-corpus sur lequel notre analyseur obtient les plus mauvais résultats sont la partie à niveau de bruit élevé des données issues du forum Doctissimo (f-mesure de 39%), et les meilleurs résultats sont ceux atteints sur la partie la moins bruitée de ces mêmes données (f-mesure de 70%). En comparaison avec les scores de 83% à 89% obtenus par différents analyseurs sur le FTB, ces résultats sont bien inférieurs. Cela montre la pertinence de développer un corpus comme le FSMB.

34. Ici comme au chapitre 9, nous utilisons la répartition standard entre sous-corpus d'entraînement, de développement et de test.

	Sous-corpus de développement					Sous-corpus de test				
	LR	LP	F1	TA	Inc.	LR	LP	F1	TA	Inc.
Doctissimo										
B+	37,22	41,20	39,11	51,72	40,47	–	–	–	–	–
B-	69,68	70,19	69,94	77,96	15,56	70,10	71,68	70,88	79,14	15,42
JeuxVideo	66,56	66,46	66,51	74,56	20,46	70,59	71,44	71,02	75,70	19,88
Twitter										
B+	62,07	64,14	63,09	64,89	31,50	54,67	58,16	56,36	64,40	32,84
B-	68,06	69,21	68,63	79,70	24,70	71,29	73,45	72,35	78,88	24,47
Facebook										
B+	–	–	–	–	–	55,26	59,23	57,18	54,64	50,40
B-	55,90	58,71	57,27	64,34	38,25	60,98	61,79	61,38	70,68	29,52
Total	*64,13*	*65,48*	*64,80*	*72,69*	*23,40*	*66,69*	*68,50*	*67,58*	*74,43*	*22,81*
FTB-uc	–	–	83,81	96,44	5,2	–	–	84,10	96,97	4,89

Tableau 9.12 – Résultats préliminaires pour l'analyse syntaxique en constituants de chaque sous-corpus du French Social Media Bank, en fonction notamment de la distinction entre sous-corpus faiblement bruités (B-) et fortement bruités (B+). Les colonnes LR, LP et F1 donnent respectivement les rappels, précisions et f-mesures obtenus. Les colonnes TA fournissent les scores d'exactitude de l'étiquetage morphosyntaxique induit par l'analyseur. Enfin, les colonnes Inc. indiquent les pourcentages de tokens inconnus de l'analyseur.

9.4.2 Analyse syntaxique de textes bruités : la campagne SANCL 2012 sur le Google Web Treebank

Nos expériences sur l'analyse statistique en constituants de corpus bruités sont parallèles à nos expériences sur l'étiquetage morphosyntaxique présentées à la section 8.4.3. Une première différence entre nos travaux en analyse syntaxique réalisés sur l'anglais dans le cadre de la campagne SANCL 2012 (Petrov et McDonald, 2012) et ceux réalisés sur le français (cf. section précédente) réside dans le fait, déjà mentionné précédemment à propos des expériences d'étiquetage morphosyntaxique, que les corpus à analyser (corpus de test) nous étaient totalement inconnus lorsque nous les avons reçus des organisateurs de la campagne SANCL 2012.

La deuxième différence est que nous avons utilisé l'analyseur syntaxique LORG [35] (Attia *et al.*, 2010), analyseur PCFG-LA que nous avons rapidement décrit à la section 9.2.3 [36].

La troisième différence, plus importante, est que nous avons cherché à améliorer nos scores en combinant nos outils de correction adaptés aux corpus bruités (cf. section 8.4)

35. Version de décembre 2012.

36. Comme à la section 9.2, les résultats présentés ici reposent sur le modèle générique pour le traitement des mots inconnus (cf. section 9.2.3). Dans les données d'entraînement, les hapax sont considérés comme inconnus.

avec deux techniques permettant de contrecarrer la dispersion des données et le changement de domaine : le clustering lexical et l'auto-apprentissage.

La technique de clustering lexical que nous avons utilisée est celle de Candito *et al.* (2011), qui s'appuie sur différents travaux publiés précédemment (Koo *et al.*, 2008 ; Candito *et al.*, 2009b ; Candito et Seddah, 2010). Cette technique permet notamment de s'adapter à un nouveau domaine, c'est-à-dire de permettre à l'analyseur de mieux traiter des données dont le domaine, le *domaine cible*, n'est pas le même que le *domaine source*, celui des données d'apprentissage (en l'espèce, la version Ontonotes 4.0 du Penn TreeBank). L'idée générale consiste à remplacer les mots par des identifiants de clusters appris sur un corpus combinant données du domaine source et du domaine cible [37]. Dans le cas de la campagne SANCL 2012, le domaine cible est en réalité multiple, puisqu'il couvre les cinq domaines indiqués ci-dessus. Pour l'apprentissage de nos clusters, nous avons utilisé les données brutes fournies par les organisateurs, dont nous avons extrait 1 000 clusters à partir des mots apparaissant au moins 100 fois. Les informations concaténées aux identifiants de clusters avant remplacement des tokens sont les trois derniers caractères du token et un marqueur booléen de capitalisation. Les identifiants de clusters ainsi étendus sont simplement appelés « clusters » dans la suite de cette section.

Nous avons complété ce clustering lexical par deux étapes distinctes d'auto-apprentissage, grâce à une première version de l'analyseur entraîné sur le sous-corpus d'entraînement d'Ontonotes (*analyseur de base*) [38]. Tout d'abord, les données annotées fournies par les organisateurs ne contenaient aucune occurrence de deux des étiquettes morphosyntaxiques prévues par le guide d'annotation du Google Web TreeBank, et deux autres étiquettes y sont très rares (fréquence inférieure à $2 \cdot 10^{-5}$). Il nous a donc fallu compléter les données d'apprentissage par de nouvelles données annotées automatiquement afin de couvrir correctement ces quatre étiquettes. Pour cela, nous avons étiqueté avec MElt l'ensemble des données brutes fournies par les organisateurs, puis nous avons sélectionné 50 phrases pour chacune de ces quatre

37. Pour cela, on applique quatre étapes successives : (i) on construit de façon non supervisée des clusters de tokens à partir d'un corpus mélangeant des données du domaine source et du domaine cible, afin de construire des clusters combinant des tokens des deux domaines (d'où le terme de « pont lexical » parfois utilisé pour dénoter cette technique) ; (ii) on entraîne l'analyseur syntaxique sur le corpus d'entraînement dans lequel on substitue à chaque token l'identifiant de son cluster, éventuellement complété par des informations supplémentaires destinées à ce que les informations morphologiques ne soient pas totalement perdues ; (iii) cette même opération de substitution est réalisée sur le texte à analyser ; (iv) une fois le texte analysé, on réintroduit les tokens d'origine à la place des identifiants de clusters.

38. La construction de l'analyseur de base se fait sur les données d'entraînement dans lesquelles les étiquettes morphosyntaxiques ont été remplacées par les étiquettes prédites par MElt, afin d'optimiser la capacité de l'analyseur à construire les bonnes structures en constituants à partir des sorties de MElt, y compris lorsque les étiquettes qu'il produit diffèrent des étiquettes de référence. À cette fin, nous appliquons la méthode du *jackknife* avec 20 tranches. C'est le cas pour tous les analyseurs dont il sera question dans cette section.

étiquettes rares ou inconnues [39]. Une fois annotées par l'analyseur de base [40], ces phrases constituent l'extension d'amorçage (ci-après AA_{Am}) à ajouter au sous-corpus d'entraînement d'Ontonotes. Par ailleurs, pour améliorer la couverture des différents domaines, nous avons sélectionné aléatoirement 70 000 phrases parmi les données brutes, avec les mêmes contraintes que pour la construction d'AA_{Am} : phrases de 7 à 20 tokens contenant au plus un couple (token, étiquette) inconnu. Une fois annotées par l'analyseur de base, ces phrases constituent l'extension générique (ci-après $AA_{Gén}$).

La table 9.13 récapitule quelques informations quantitatives sur différentes variantes du corpus d'entraînement.

	Ontonotes-train + AA_{Am}		Ontonotes-train + AA_{Am} + $AA_{Gén}$
Unités élémentaires	tokens	clusters	clusters
Taille du vocabulaire	36 052	15 073	19 893
Nombre d'unités ($\times 10^6$)	0,7		1,66
Nombre de phrases	30 220		99 433

Tableau 9.13 – Informations quantitatives sur les corpus d'entraînement en fonction du type d'unités élémentaires (tokens ou clusters). Les corpus d'entraînement sont composés du sous-corpus d'entraînement du corpus Ontonotes complété par les données d'auto-apprentissage : l'extension d'amorçage (AA_{Am}) et éventuellement l'extension générique ($AA_{Gén}$).

L'analyseur syntaxique LORG permet de choisir le nombre x de cycles *split/merge* appliqués lors de la construction de l'analyseur syntaxique PCFG-LA — cf. Matsuzaki *et al.* (2005) et Petrov *et al.* (2006) pour plus de détails —, mais également de s'appuyer sur le produit de y grammaires distinctes afin de réduire la part d'aléatoire de l'algorithme d'apprentissage, ce qui permet d'améliorer les performances (Petrov, 2010). Dans la suite, un analyseur appris avec une certaine valeur de x et une certaine valeur de y sera considéré comme relevant d'une configuration notée s.*x*n*y*.

Au moment d'analyser des données nouvelles, l'étiquetage morphosyntaxique est réalisé de l'une ou l'autre des deux façons suivantes : soit il est réalisé intégralement par l'analyseur syntaxique (pas d'étiquetage *a priori*), soit MElt est utilisé pour étiqueter les données mais ne sont conservées et fournies à l'analyseur que les étiquettes associées à des tokens tels que le couple (token, étiquette) est inconnu des données d'apprentissage (soit le token est inconnu, soit il n'a jamais été vu avec cette étiquette) [41].

39. La sélection a été aléatoire parmi les phrases de longueur moyenne (7 à 20 mots) et ne contenait aucun autre couple (token, étiquette) inconnu du sous-corpus d'entraînement d'Ontonotes que celui contenant l'étiquette rare ou inconnue. Ces restrictions ont naturellement pour objectif de sélectionner des phrases aussi faciles à analyser automatiquement que possible.

40. L'analyseur de base ignore les étiquettes qui lui sont inconnues, et en met d'autres à la place. Dans ce cas, les étiquettes produites par MElt sont rétablies dans les arbres produits par l'analyseur de base.

41. Dans ce cas, en cas d'échec de l'analyseur, ce dernier est relancé sans que ne lui soit plus fourni d'étiquette.

Nous avons obtenu nos premiers résultats, avant la campagne proprement dite, sur les données de développement. Ces résultats, fournis à la table 9.14, sont déjà élevés. Ils montrent l'impact positif, mais non massif, qu'il y a à utiliser MElt et à fournir à l'analyseur les étiquettes des couples (token, étiquette) inconnus. Des expériences où nous fournissions toutes les étiquettes produites par MElt ont en revanche conduit à des résultats inférieurs.

On peut donc se demander si le débruitage et l'étiquetage produits par MElt sont vraiment utiles, en particulier dans un système utilisant le clustering lexical pour réduire le taux d'inconnus et la dispersion des données : corriger un token altéré ou le mettre dans le même cluster que sa version non altérée peuvent être vus comme deux moyens différents de pallier le même problème. Il n'en reste pas moins que la gestion des étiquettes spécifiques aux données issues du web ne peut, par définition, se faire par apprentissage sur un corpus qui n'en contient pas. De plus, les corrections réalisées par l'outil de débruitage intégré à MElt ne peut qu'améliorer la qualité des clusters appris : ces deux approches sont complémentaires bien plus que redondantes. C'est ce qui explique que nos meilleurs résultats, en moyenne, soient obtenus avec MElt.

	Ontonotes (dev)	E-mails	Blogs
sans clustering, sans étiquetage préalable par MElt			
Score d'analyse syntaxique (f-mesure `evalb`$^+$)	89,96	80,05	84,71
Précision de l'étiquetage morphosyntaxique induit	95,58	86,12	92,24
avec clustering, sans étiquetage préalable par MElt			
Score d'analyse syntaxique (f-mesure `evalb`$^+$)	90,09	80,61	**85,43**
Précision de l'étiquetage morphosyntaxique induit	96,32	88,23	94,02
avec clustering, étiquetage préalable par MElt seulement sur les couples (token, étiquette) inconnus			
Score d'analyse syntaxique (f-mesure `evalb`$^+$)	**90,17**	**81,06**	85,36
Précision de l'étiquetage morphosyntaxique induit	96,60	90,81	94,80

Tableau 9.14 – Résultats de l'analyseur de base sur les sous-corpus de développement, en configuration s5+n4 sans auto-apprentissage. Les scores d'analyse syntaxique sont des f-mesures produites par l'outil `evalb`$^+$ (cf. section 9.4.1).

Les résultats obtenus sur les corpus de test, y compris les résultats officiels, sont fournis dans la table 9.15 [42] Bien que nous n'ayons pu fournir dans les temps nos meilleurs

42. On constate que les résultats sont meilleurs avec les configurations de type s5 qu'avec celles de type s6, qui souffrent vraisemblablement de sur-apprentissage, notamment au vu des résultats de la configuration s6n8, meilleurs que ceux de la configuration s5n8 sur Ontonotes mais moins bons sur les données issues du web. Un tel résultat peut sembler surprenant, par exemple au vu des conclusions de Huang et Sagae (2010) qui ont obtenu une dégradation de leurs scores à partir de 170 000 phrases d'auto-apprentissage et avec des configurations de type s7. L'explication tient vraisemblablement à la prudence de nos critères de sélection des phrases ajoutées au corpus d'entraînement, critères qui entraînent une certaine homogénéité de ces phrases mais également une relative similarité avec le domaine de départ, celui d'Ontonotes (rappelons que les phrases retenues ne contenaient qu'au plus un inconnu).

résultats (ceux des configurations s5), nos deux configurations s6 sont arrivées deuxième et troisième de la campagne, derrière les résultats produits par Le Roux *et al.* (2012a) [43] mais devant les équipes de Stanford et d'IMS, pourtant régulièrement très bien classés dans les campagnes d'évaluation des analyseurs syntaxiques.

	BKY (baseline)	Alpage (off.)		DCU-P13	Alpage (non off.)	
Configuration		s6n4	s6n8		s5n4	s5n8
Rang		3	2	1	(2')	(1')
Questions	*75,92*	80,52	**80,60**	*82,19*	81,37	**81,46**
Newsgroups	*78,14*	83,67	**84,03**	*84,33*	83,84	**84,13**
Recensions	*77,16*	81,52	**81,76**	*84,03*	82,55	**82,68**
Total	*77,07*	81,90	**82,13**	*83,52*	82,34	**82,45**
WSJ	*88,21*	**89,91**	89,87	*90,53*	89,60	89,74

Tableau 9.15 – Résultats de l'analyseur avec auto-apprentissage sur les trois sous-corpus de test du Google Web Treebank et sur la section de test d'Ontonotes. Nos résultats officiels pour la campagne SANCL 2012 sont indiqués sous « Alpage (off.) ». Les deux configurations concernées, de type s6+n4 et s6+n8, sont arrivées deuxième et troisième. Les résultats des vainqueurs (Le Roux *et al.*, 2012a) sont donnés dans la colonne DCU-P13. Les résultats de notre analyseur en configuration s5+n4 et s5+n8 sont également indiqués. Sur les données du Google Web Treebank ils sont meilleurs que nos résultats officiels mais n'ont pas pu être obtenus à temps pour être soumis officiellement.

9.5 Éléments de conclusion

Les expériences relatées dans ce chapitre montrent le bénéfice que l'on peut tirer de l'utilisation d'informations lexicales morphologiques et/ou syntaxiques, de façon directe ou indirecte, dans différents types d'analyseurs syntaxiques, en l'espèce symboliques, statistiques et hybrides, pour produire différents types de structures syntaxiques (constituants, dépendants, analyses TAG).

L'exploitation d'informations lexico-syntaxiques dans les analyseurs syntaxiques reste néanmoins un problème ouvert. En particulier, nous ne sommes pas au courant, aujourd'hui, de travaux ayant cherché à exploiter de telles informations (valence lexicale, notamment) dans des analyseurs syntaxiques neuronaux. Une piste dans cette direction est en cours de développement par É. de La Clergerie, qui consiste à coupler l'analyseur frmg, qui s'appuie sur une grammaire et sur les informations lexicales morphologiques et syntaxiques du Le*fff*, avec une architecture de désambiguïsation qui ne serait plus

43. La campagne SANCL 2012 permettait également de soumettre des résultats d'analyse syntaxique en dépendances. Nous avons cependant choisi de ne produire que des résultats en constituants.

statistique mais neuronale, voire qui couplerait techniques statistiques et neuronales. Il s'agirait là d'une double hybridation combinant donc approches symboliques, statistiques et neuronales. Le couplage statistique–neuronal a du reste déjà été mis en œuvre dans certains des analyseurs que nous avons déployés pour notre participation à la campagne d'évaluation UD 2017 évoquée au chapitre précédent (Villemonte de La Clergerie *et al.*, 2017) [44]. Mais le double couplage symbolique–statistique–neuronal, l'une des voies qui permettront de tirer parti, dans une même architecture d'analyse syntaxique, de la puissance des réseaux de neurones et des informations fournies par un lexique comme le Le*fff*, est encore à ce jour un programme de recherche à défricher.

Il en est de même pour des architectures neuronales qui seraient capables d'exploiter directement les informations lexico-syntaxiques fournies par un lexique comme le Le*fff*, à l'image de ce que nous avons réalisé au niveau morphosyntaxique avec alNNtagger (cf. section 8.3). Mais la situation est bien plus délicate au niveau syntaxique, les informations à prendre en compte étant plus complexes et plus structurées, et les structures à produire étant bien plus complexes qu'une simple séquence d'étiquettes. Une piste possible, sur laquelle je prévois à l'avenir des travaux en collaboration, notamment, avec É. de La Clergerie, consisterait à utiliser une architecture neuronale pour désambiguïser les forêts d'analyse produites par l'analyseur FRMG, qui s'appuie notamment sur le Le*fff*. L'idée serait d'entraîner un analyseur neuronal par transitions de type « shift-reduce » (Nivre *et al.*, 2007b) qui serait guidé par la forêt produite par FRMG. La difficulté est que FRMG produit des analyses qui suivent ses propres conventions, lesquelles diffèrent des schéma d'annotation utilisés par les corpus arborés disponibles. Demander à ce ré-analyseur neuronal de désambiguïser la forêt d'analyse en tant que telle nécessiterait de développer un oracle (statique ou dynamique) à partir de corpus arborés aux annotations hétérogènes par rapport aux forêts produites par FRMG, ce qui est délicat. Il est donc plus adapté de faire en sorte que ce ré-analyseur produise directement des analyses qui suivraient les conventions de tel ou tel corpus arboré. Ainsi, il pourraît être entraîné directement à partir d'un corpus arboré utilisé comme entraînement, la forêt d'analyse produite pour chaque phrase d'entraînement par FRMG étant utilisée comme source d'informations pour le réseau de neurones. L'avantage d'une telle approche est que l'on pourrait utiliser une architecture neuronale multi-tâche, où chaque tâche correspond à un schéma d'annotation particulier. Un système unique pourrait ainsi être entraîné sur tous les corpus arborés disponibles pour le français. Il permettrait de produire pour une phrase donnée en entrée les analyses correspondant à chacun des guides d'annotation correspondant. Ce serait là un moyen direct d'exploiter au mieux

44. La raison pour laquelle nous n'avons pas insisté dans ce document sur ces analyseurs syntaxiques est que notre contribution concrète à notre participation à la campagne UD 2017 s'est porté avant tout, comme décrit au chapitre précédent, sur les aspects de tokenisation, segmentation en phrases et surtout d'analyse morphosyntaxique. Le développement des analyseurs syntaxiques proprement dits, et notamment le développement d'analyseurs hybrides statistiques–neuronaux, a été avant tout le fait d'É. de La Clergerie.

les informations de la métagrammaire de FRMG et du Le*fff* tout en s'adaptant facilement à différentes conventions d'annotation, couplant ainsi description linguistique riche et architecture neuronale.

Chapitre 10

Conclusion et perspectives

10.1 Conclusion

Les travaux présentés dans ce document donnent un aperçu de la diversité des travaux en traitement automatique des langues et en linguistique computationnelle en rapport avec le lexique. Ils couvrent des aspects liés à la modélisation des informations lexicales, au développement de ressources lexicales et à l'exploitation de telles ressources dans des systèmes de traitement automatique. Ils ont conduit au développement de plusieurs ressources lexicales relevant des niveaux morphologique, syntaxique et sémantique, et ce pour plus d'une dizaine de langues typologiquement variées, avec néanmoins une prédilection pour le français.

Comme il est apparent au travers de la liste de mes co-auteurs, je n'ai pas réalisé ces travaux sans interaction avec mon environnement scientifique. Mes collaborations et encadrements de thèses et de post-docs au sein de l'équipe ALPAGE puis ALMAnaCH (cf. introduction), au sein du LabEx *Empirical Foundations of Linguistics*, au sein des différents projets dans lesquels j'ai été impliqué, mais aussi au fil des rencontres et des opportunités m'ont permis de découvrir des champs d'étude, des ressources lexicales, des langues, des idées et des problématiques que je n'aurais peut-être jamais rencontrées autrement.

Pour autant, les travaux présentés dans ce document n'apportent pas, tant s'en faut, des réponses à toutes les problématiques liées au lexique. La révolution statistique puis, plus encore peut-être, la révolution neuronale ont bouleversé le domaine du traitement automatique des langues, et nécessitent de renouveller notre compréhension du rôle des ressources lexicales. Mais de telles recherches ne peuvent s'inscrire que dans un programme de recherche plus large qui mette en musique le traitement automatique des langues et la linguistique computationnelle, mais également les humanités numériques, dès lors qu'elles constituent à la fois une source d'informations linguistiques, notamment

pour les langues anciennes, et un domaine d'application naturel et enrichissant pour le traitement automatique des langues. C'est dans cet esprit que j'ai coordonné la construction du programme scientifique de la nouvelle équipe ALMAnaCH dont je suis le responsable, programme qui constituera le cadre naturel de mes recherches futures. La section suivante décrit brièvement ce programme de travail.

10.2 Programme scientifique de l'équipe ALMAnaCH

L'équipe ALMAnaCH est une équipe pluri-disciplinaire, à l'interface entre informatique, linguistique, philologie et statistique, qui rassemble traitement automatique des langues, linguistique computationnelle et humanités computationnelles [1]. Elle conserve de l'équipe ALPAGE l'objectif de développer des systèmes de traitement automatique des langues au niveau de l'état de l'art qui puissent être utilisés à la fois dans le milieu académique et par des industriels, de les mettre en œuvre dans des systèmes d'extraction d'informations et de fouille de textes, et d'étudier la modélisation des langues pour en améliorer la compréhension. Mais ces objectifs seront renforcés par l'intégration de thématiques de recherche liées aux humanités computationnelles, et notamment par l'apport de recherches sur la modélisation de l'évolution des langues et, par conséquent, sur les langues anciennes. Cette ouverture thématique a motivé la mise en place d'une collaboration nouvelle, avec l'École Pratique des Hautes Études (ÉPHÉ), établissement d'excellence dans plusieurs domaines dont les sciences historiques et philologiques, y compris au moyen d'approches computationnelles.

L'un des enjeux les plus importants en traitement automatique des langues est la modélisation et la prise en compte de la variation linguistique. Les données textuelles varient selon différents facteurs :

— le domaine et le genre des textes (dépêches d'agence, littérature scientifique, poésie, transcription de données orales…), avec leurs spécificités terminologiques mais aussi syntaxiques et stylistiques (textes juridiques) ;

— le niveau de complexité linguistique, par exemple lorsque l'on s'adresse à des personnes en situation de handicap [2] ;

— les variables sociodémographiques (âge, milieu social, niveau d'éducation) ; cette variation est notamment attestée sur les médias sociaux, et se manifeste en

1. J'utilise à dessein le terme d'*humanités computationnelles* plutôt que le terme plus répandu d'*humanités numériques* car je fais une différence entre les deux : les humanités numériques utilisent l'informatique comme une source de techniques et de technologies pour explorer des questions de recherche en sciences sociales et humanités (SHS), alors que les humanités computationnelles ont pour ambition d'améliorer l'état de l'art à la fois en informatique et en SHS, en impliquant l'informatique comme un vrai domaine de recherche. À cet égard, la linguistique computationnelle et le traitement automatique des langues peuvent être considérés comme les domaines de recherche les plus anciens relevant des humanités computationnelles.

2. Ainsi, une collaboration est en cours de démarrage avec Facebook, l'UNAPEI et le Secrétariat d'État chargé des personnes handicapées, pour les personnes souffrant de handicap mental.

particulier par des spécificités orthographiques, lexicales et syntaxiques (langue non canonique) ;

— l'évolution permanente des langues à toutes les échelles de temps : changements phonétiques, lexicaux (emprunts, création lexicale…), morphologiques, syntaxiques, etc. ; de plus les textes anciens ne suivent généralement pas une orthographe normée ;

— les erreurs introduites par les systèmes d'OCR (reconnaissance optique de caractère) et d'HTR (transcription automatique de textes manuscrits) ; cela ne relève pas à proprement parler de la variation linguistique, mais il y a de nombreux points communs avec les variations liées aux facteurs sociodémographiques mentionnés ci-dessus ainsi qu'avec les flottements orthographiques que l'on observe dans certains types de textes anciens (en ancien et moyen français, par exemple).

Maîtriser cette variabilité multiforme est, encore aujourd'hui, un problème ouvert en traitement automatique des langues. Les approches couramment utilisées, qui s'appuient souvent sur des techniques d'apprentissage automatique supervisées ou semi-supervisées, nécessitent que des volumes importants de données annotées soient disponibles. Elles peinent encore à traiter correctement les niveaux élevés de variabilité que l'on rencontre par exemple dans les textes produits par les utilisateurs sur les médias sociaux ou dans les textes anciens reflétant des états antérieurs de la langue.

Nous nous attaquerons à l'enjeu de la variabilité linguistique selon deux dimensions complémentaires.

L'une de ces directions consiste à améliorer notre capacité à produire des représentations linguistiques à des niveaux qui sont moins affectés par la variation linguistique. Cela nécessitera tout d'abord d'améliorer l'état de l'art de l'analyse sémantique des textes, et pas seulement de leur analyse syntaxique ou morphosyntaxique. Cela passera également par l'intégration aux systèmes de traitement automatique des langues d'informations contextuelles, qu'elles soient linguistiques (phrases précédentes…) ou non-linguistiques pour améliorer l'analyse. Il s'agit là d'un champ de recherche émergent et prometteur. Il nous faudra donc identifier, modéliser et exploiter les différents types d'informations contextuelles pertinentes. Il sera alors possible, par exemple, d'améliorer la qualité des systèmes d'extraction d'informations et de connaissances, notamment dans des documents relevant de domaines techniques et dans des documents anciens à valeur historique, travaux qui relèvent des humanités numériques et computationnelles. Mais la prise en compte du contexte nous permettra également d'initier des recherches autour des contenus conversationnels. Nous travaillerons ainsi sur applications telles que les systèmes de *chatbots*.

La seconde direction de recherche en rapport avec la variabilité linguistique est celle de sa modélisation et de sa meilleure compréhension. À cet égard, nous nous concentrerons

en particulier sur trois types de variation linguistique, tous mentionnés plus haut : la variation d'origine sociodémographique, la variation en termes de complexité et la variation diachronique. Mais les points communs entre les différents types de variations suggèrent de travailler de façon aussi généraliste que possible à leur prise en compte, notamment au niveau de la variation surfacique (orthographe non canonique ou non normée, erreurs d'OCR et d'HTR) et lexicale (mots anciens, mots techniques, néologismes).

Ces deux directions de recherche reposent sur la disponibilité de ressources linguistiques (corpus, lexiques). Je participerai donc au développement de corpus bruts à partir de sources originales, au développement de nouveaux corpus annotés, mais également au développement de ressources lexicales. En cohérence avec ce que j'indique ci-dessous, je porterai une attention particulière, mais non exclusive, au développement de ressources pour les langues anciennes, y compris pour l'ancien et le moyen français.

L'ensemble de ces travaux s'appuiera de façon jointe sur des compétences linguistiques solides et sur le développement d'approches informatiques adaptées (formelles, statistiques et neuronales, et hybrides lorsque cela est pertinent).

Au sein de ce programme de recherche, il y a une thématique qui m'intéresse plus particulièrement. L'ensemble des travaux présentés dans ce document relève d'une étude synchronique des langues. Or comprendre l'histoire des langues au fil des siècles et des millénaires est un moyen unique d'aborder les principes sous-jacents à la structuration des systèmes linguistiques et à leur évolution, et ce à tous les niveaux d'analyse. C'est également un moyen de reconstituer, ne serait-ce que partiellement et imparfaitement, ce qu'étaient les langues et les cultures de nos ancêtres. Pourtant, peu de travaux ont réellement cherché à mettre les avancées des techniques de traitement automatique des langues, et notamment celles liées au développement de ressources lexicales, au service de la linguistique historique et de l'étymologie. C'est là mon ambition pour les prochaines années, et la dernière section de ce dernier chapitre esquisse quelques pistes que je souhaite poursuivre dans cette direction.

10.3 Vers une étymologie computationnelle

Les approches quantitatives et computationnelles ont joué un rôle de premier plan dans pratiquement tous les domaines de la linguistique, y compris en phonologie, en morphologie et en lexicologie. La linguistique de corpus assistée par ordinateur est apparue dans les années 1970, suivie par de nouvelles directions de recherche telles que l'étude computationnelle de la complexité linguistique et l'implémentation informatique de modèles morphologiques, syntaxiques et sémantiques. Cette interaction entre la linguistique, l'informatique puis les statistiques a entraîné une évolution majeure dans

la plupart des domaines relevant de la linguistique, qui bénéficient de plus en plus de nouvelles informations empiriques sur la langue et les langues, ainsi que d'approches reproductibles et falsifiables.

Cependant, ce changement de paradigme n'a pas encore bénéficié aux études étymologiques, et plus généralement à la linguistique historique et comparative. L'une des principales raisons en est le nombre relativement faible de contacts et donc de collaborations entre, d'un côté, les linguistes historiques, les comparatistes et les étymologistes et, d'un autre côté, les informaticiens travaillant sur les données linguistiques.

Et pourtant, les efforts pour mettre en œuvre des approches quantitatives pour étudier des problématiques de linguistique historique remontent aux années 1950, en particulier grâce au travail de Swadesh. Son point de départ consistait en un inventaire réduit de concepts de base (100 concepts dans Swadesh (1971)) conçus pour être tout à la fois universels (réalisés dans toutes les langues) et stables (peu sujets à l'innovation lexicale). L'hypothèse de Swadesh était que le niveau de proximité entre deux langues pouvait être induit à partir du nombre de ces concepts de base dont les traductions dans deux langues étaient génétiquement apparentées, tel que reconnues par des experts. Swadesh et d'autres ont utilisé cette métrique pour détecter les relations entre langues et pour créer des arbres retraçant l'histoire des langues et de leur divergence, également appelés phylogénies linguistiques. Cette approche, appelée *lexicostatistique*, a été étendue par l'hypothèse que les mots des listes de Swadesh étaient remplacés à un taux fixe au cours de l'histoire des langues. Si l'on est capable d'estimer ce taux [3], tout nœud dans un arbre phylogénétique donné peut être daté, y compris la racine de l'arbre, qui représente l'ancêtre commun de toutes les langues considérées. Cette méthode, appelée *glottochronologie*, a inspiré de nombreux algorithmes et outils plus récents de construction automatique d'arbres phylogénétiques. En particulier, les jugements d'experts utilisés pour la constitution des bases de données de cognats qui servent d'entrée à de telles méthodologies ont été parfois remplacés par des algorithmes automatiques de détection de cognats (Kondrak, 2009 ; List *et al.*, 2017 ; Jäger *et al.*, 2017). Mais les algorithmes de création et de datation d'arbres phylogénétiques produisent toujours des résultats instables et peu fiables. Ainsi, plusieurs modèles récents (Gray et Atkinson, 2003 ; Ryder et Nicholls, 2011 ; Bouckaert *et al.*, 2012) font remonter le proto-indo-européen, ancêtre commun de langues telles que les langues germaniques, celtiques, romanes, grecques et indo-iraniennes, à 8000–9000 ans avant notre ère, une date désormais universellement reconnue comme étant d'environ 5000 ans trop ancienne [4].

3. Sur les 100 concepts de la liste de Swadesh, un taux de 14 remplacements par millénaire a été estimé.

4. Même Colin Renfrew, qui était l'un des plus ardents défenseurs de cette datation précoce, a aujourd'hui changé d'avis.

Les algorithmes de détection de cognats et les algorithmes d'alignement phonétique associés (Prokić et Cysouw, 2013 ; List, 2014) commencent aujourd'hui à apporter une aide concrète à la recherche en linguistique comparative, en particulier pour les familles linguistiques les moins étudiées (Hill et List, 2017). Néanmoins, une compréhension en profondeur de l'histoire des langues ne peut être atteinte qu'en comprenant l'étymologie des mots qui les constituent, ce qui commence par l'identification des cognats et des correspondances phonétiques systématiques entre langues, mais va naturellement bien au-delà. Ce sont ces thématiques dont le renouvellement sera central dans mes recherches futures. Les questions et problèmes qui s'y posent sont multiples. Premièrement, la quasi-totalité des algorithmes de détection de cognats repose sur des comparaisons formelles de surface dans le but de détecter les cognats directs, c'est-à-dire des mots directement hérités d'un même étymon dans la langue parente. Certains de ces outils reposent sur des algorithmes d'alignement au niveau phonétique, mais d'autres s'appuient sur des jeux de données qui ne sont même pas phonétisés et parfois d'une qualité insuffisante (Hauer et Kondrak, 2011). Deuxièmement, l'héritage direct à partir d'un ancêtre commun est loin d'être le seul type d'étymologie. De nombreux mots sont créés en synchronie, au moyen de mécanismes de création lexicale. Ces mécanismes comprennent des processus morphologiques internes à la langue tels que la dérivation affixale, la composition et la création analogique (par exemple, la dérivation inverse, cf. Garnier, 2016), ainsi que des processus inter-langues, dans le cas des emprunts. De plus, les mécanismes de réfection analogiques peuvent affecter l'histoire de la morphologie d'un mot de multiples manières, par exemple par la restructuration de son paradigme morphologique ou même le changement de sa classe flexionnelle (McCarthy, 2005 ; Albright, 2008 ; Garrett, 2008). Pour finir, le sens d'un mot peut changer au fil du temps, le plus souvent selon des tendances non triviales qui obscurcissent l'identification de cognats.

Quelques travaux préliminaires récents explorent les évolutions sémantiques au moyen d'approches quantitatives (Dellert, 2016). Cependant, aucun modèle computationnel actuel de l'évolution des langues ne tient compte du rôle-clef de la morphologie, tant flexionnelle que dérivationnelle, dans l'étymologie. Même la reconstruction d'étymons à partir de cognats, une tâche clé pour les comparatistes, a rarement été étudiée au moyen de méthodes computationnelles (voir toutefois les résultats préliminaires de Bouchard-Côté *et al.*, 2013).

Il n'est donc pas surprenant que la recherche en étymologie soit encore exclusivement réalisée par des experts humains qui suivent la méthode comparative, laquelle s'appuie sur une connaissance approfondie des langues modernes et anciennes et de leurs attestations épigraphiques, paléographiques et littéraires. L'objectif de cette méthode est de construire des modèles détaillés de l'histoire des langues étudiées et de l'étymologie des mots constituant leurs lexiques. Le principe de base de la méthode comparative est l'hypothèse

que le changement phonétique est systématique et régulier : toutes les occurrences d'un environnement phonétique pertinent dans le lexique d'une langue subiront le même changement phonétique. Outre la phonétique, les étymologistes prennent également en compte tous les niveaux de l'évolution des langues, y compris la sémantique et, surtout, la morphologie.

Le but de mes recherches sera la mise en œuvre simultanée d'algorithmes informatiques prenant en compte les niveaux phonétique, morphologique et sémantique, ouvrant ainsi ainsi de nouvelles voies de recherche. Des bases de données lexicales à grande échelle couvrant de nombreuses langues et s'appuyant sur les connaissances approfondies d'experts du domaine, couplées à des modèles computationnels des changements phonétiques et, surtout, morphologiques, permettraient de :

— vérifier et affiner des hypothèses existantes, par exemple en aidant les comparatistes à améliorer la cohérence et la chronologie relative des lois phonétiques qu'ils proposent ;

— proposer automatiquement de nouvelles étymologies et de nouvelles hypothèses comparatives ;

— mieux comprendre les causes de certains changements, notamment au niveaux morphologique et sémantique ;

— valider ou même créer des phylogénies qui s'appuient sur des inventaires d'étymologies plausibles.

J'aimerais ainsi développer et utiliser des approches computationnelles pour étudier le changement linguistique au fil du temps, en mettant l'accent sur les langues indo-européennes en général et le français en particulier. Dans la continuité des travaux présentés dans ce document, je me concentrerai notamment sur l'évolution du lexique. Je prendrai en compte non seulement les changements phonétiques mais également les évolutions au niveau morphologique, dans la suite de certains des travaux présentés aux chapitres 2 et 4, et celles au niveau sémantique, peut-être en m'appuyant sur des ressources de type wordnet (cf. chapitre 6). Cela passera par le développement de lexiques morphologiques pour un certain nombre de langues, y compris par l'extraction d'informations linguistiques à partir de ressources semi-structurées comme des dictionnaires électroniques. Il me faudra également trouver des réponses à des problèmes informatiques complexes, notamment formels et algorithmiques, tout en étant en prise directe avec des questions de modélisation de l'histoire des langues et de reconstruction de proto-langues.

Je compte pour cela collaborer avec des linguistes computationnels, des linguistes comparatifs et historiques et des linguistes descriptifs, afin de développer ensemble des ressources, des modèles et des outils qui permettront d'aborder d'une façon renouvelée les recherches sur l'histoire des mots. Une telle entreprise me permettra de mettre en

œuvre dans un contexte nouveau un certain nombre de compétences que j'ai acquises au fil des années et que j'ai évoquées dans ce document, notamment pour le développement de lexiques morphologiques et sémantiques et en morphologie formelle et quantitative. Elle me permettra également d'approfondir de nouveaux sujet, notamment en linguistique historique et en diachronie.

Annexes

Annexe A

Panorama historique et thématique du traitement automatique des langues

A.1 Aperçu des approches formelles de la morphologie [1]

Les approches contemporaines de la morphologie proposent des modélisations divergentes de la formation des mots. Deux des principales oppositions entre les différents modèles proposés correspondent à des réponses différentes à deux questions fondamentales que nous allons analyser successivement : (i) morphologie et syntaxe sont-elles autonomes l'une par rapport à l'autre, ou s'agit-il d'une distinction traditionnelle entre phénomènes qu'il convient de modéliser de façon unifiée et de regrouper ainsi en une notion étendue de ce qu'est la syntaxe ? (ii) l'unité élémentaire, dans un lexique morphologique, est-elle le mot (nous laissons sous-spécifié le type de mot dont il s'agit eu égard aux discussions du chapitre 1) ou des unités plus petites que le mot, telles que le morphème, notion traditionnelle définie depuis Baudouin de Courtenay (1895) comme une séquence phonologique représentant une unité minimale de sens ?

A.1.1 Approches lexicales : morphologie morphématique et Morphologie Distribuée

Une approche comme celle de Lieber (1981) définit le lexique comme une collection de morphèmes, destinés à être insérés directement dans la structure syntaxique : il n'y a pas alors de morphologie indépendante de la syntaxe, et la syntaxe produit des structures de surfaces (« mots » et énoncés) de façon indifférenciée. Ce processus de construction est qualifié d'*incrémental*, puisque chacune des opérations syntaxiques à l'œuvre dans la construction d'un énoncé, y compris celles que

1. Cette section s'appuie en partie sur le mémoire de thèse de Géraldine Walther et sur le mémoire d'habilitation à diriger les recherches d'Olivier Bonami, toute inexactitude restant naturellement de mon fait.

l'on qualifie traditionnellement d'opérations morphologiques, contribue à construire incrémentalement et compositionnellement le sens de cet énoncé en parallèle à sa réalisation phonologique.

La Morphologie Distribuée (Halle et Marantz, 1993), modèle standard en linguistique chomskienne contemporaine, partage avec ce modèle une conception du lexique comme inventoriant des unités plus petites que le mot dont il s'agit de modéliser la combinaison : ces deux modèles, pour cette raison, sont qualifés de *lexicaux* par Stump (2001). En Morphologie Distribuée, toutefois, le lexique n'est pas composé de morphèmes au sens classique mais d'unités sous-spécifiées phonologiquement représentant des structures de traits combinant informations lexicales et grammaticales, unités appelées malgré tout *morphèmes*. Ces unités sont ici encore combinées par des mécanismes syntaxiques, avant que la composante phonologique ne spécifie comment construire les formes de surface (il s'agit de la *late insertion* ou insertion retardée). Une composante dédiée, appelée « encyclopédie », est en charge de gérer les spécificités lexicales. On est donc ici en présence d'un modèle qui n'est plus incrémental, la forme phonologique étant construite *in fine*. Stump (2001) parle ici de modèle *réalisationnel*, ce qui exprime le fait que la forme (phonologique) de surface est construite à partir de l'ensemble des traits à exprimer.

Ces deux modèles ont ainsi en commun de ne pas distinguer morphologie et syntaxe. Chomsky (1970) avait déjà identifié le caractère partiellement non-prédictible des changements sémantiques qui accompagnent les relations morphologiques dérivationnelles : un même mécanisme dérivationnel — dans son cas, certains mécanismes de nominalisation déverbale en anglais — peut donner, à partir de différents lexèmes donnés en entrée, des lexèmes dérivés dont le lien sémantique avec le lexème de départ varie, tant sur le plan du sens intrinsèque que de la structure actancielle. Cela fait de la morphologie dérivationnelle une opération éminemment lexicale. Il explicite ainsi, du moins à cette époque, la nécessité de traiter séparément d'une part la syntaxe et ses transformations régulières et d'autre part la morphologie dérivationnelle, à traiter dans le lexique. C'est le point de départ des approches *lexicalistes* contemporaines. On notera que ce type d'argument ne vaut que pour la morphologie dérivationnelle, la flexion étant sémantiquement régulière (cf. notamment Wurzel, 1984 ; Boyé, 2011).

Une étape supplémentaire est franchie par Bresnan (1982) qui montre qu'il existe des exemples d'opérations morphologiques dérivationnelles qui prennent en entrée non pas des lexèmes mais des formes fléchies spécifiques. La morphologie flexionnelle doit donc venir, en un sens, en amont de la morphologie dérivationnelle et non en aval comme la syntaxe. Autrement dit, les conclusions auxquelles Chomsky (1970) arrive concernant la morphologie dérivationnelle doivent être étendues à la morphologie flexionnelle. Ainsi, sous cette analyse, la syntaxe ne manipule plus que des formes fléchies : on parle de *lexicalisme fort*. La morphologie (flexionnelle et dérivationnelle) doit alors être considérée

comme autonome par rapport à la syntaxe (Zwicky et Pullum, 1988). Si, de plus, on postule une identité entre mots morphologiques (*output* de la morphologie flexionnelle) et mots syntaxiques (*atomes syntaxiques*), on en arrive au *principe d'intégrité lexicale* (Bresnan, 1982 ; Bresnan et Mchombo, 1995) selon lequel les mots construits par la morphologie sont atomiques en syntaxe et que leur structure interne est donc inaccessible à la syntaxe. Toutefois, comme évoqué au chapitre 1, il existe des contre-exemples à ce principe, que l'on pourrait analyser en termes de décalage entre mots morphologiques et mots syntaxiques.

Un autre argument en faveur de l'autonomie de la morphologie, argument de nature différente, réside dans le fait, étudié notamment par Aronoff (1994), qu'il existe des généralisations au niveau morphologique qui ne peuvent être motivées par des considérations syntaxiques ou phonologiques. C'est ce qu'il appelle la morphologie pure (*morphology by itself*). Il introduit ainsi un niveau dit *morphomique*. C'est à ce niveau que l'on peut décrire des généralisations internes à un système morphologique, et notamment des situations où, pour un nombre significatif de lexèmes, un même sous-ensemble de leurs formes fléchies partage une propriété morphologique commune alors qu'elles ne partagent aucune propriété syntaxique ou phonologique qui en constituerait une justification. Aronoff (1994) indique notamment deux exemples. Le premier est donné par les participes passés et passifs en anglais, qui sont systématiquement identiques quand bien même ils ne partagent pas davantage de propriétés syntaxiques communes entre eux qu'avec d'autres formes. Le second exemple est donné par les radicaux verbaux en latin : les formes fléchies sont formés à partir de trois radicaux distincts qui sont répartis de la même façon quel que soit le lexème verbal mais qui ne servent pas à construire des sous-ensembles de formes ayant des propriétés syntaxiques communes. Par exemple, le radical traditionnellement qualifié de radical du supin est utilisé pour construire les formes du participe passé passif, celles du participe futur actif et les formes finies perfectives du passif, et ce, quel que soit le lexème verbal considéré [2]. Un autre exemple de telle structure morphomique est donné par Corbett (2010) : en dhaasanac (couchitique, afro-asiatique), comme illustré par la table A.1, les paradigmes verbaux utilisent au présent deux formes distinctes seulement qui correspondent aux mêmes sous-parties des paradigmes, indépendamment de la façon dont elles sont construites, ces sous-parties de paradigmes n'ayant aucune propriété syntaxique homogène qui en justifie l'extension.

2. Pour peu, naturellement, que ledit lexème ait des formes passives. On notera qu'une telle description des faits repose sur l'hypothèse habituelle que le passif en latin est de nature flexionnelle et non pas dérivationnelle (Flobert, 1967 ; Baerman, 2007). Cette hypothèse est critiquée par exemple par Walther (2011a, 2013b), entre autres sur la base de la non-systématicité du lien sémantique entre formes actives et formes passives des verbes qui disposent des deux ensembles de formes, phénomène typique de la dérivation par opposition à la flexion (cf. notamment Wurzel, 1984).

	SEÐ 'marcher'		YUUFUMI 'tousser'	
	SG	PL	SG	PL
1.INCL	—	*seð*	—	*yuufumi*
1.EXCL	*seð*	*sieti*	*yuufumi*	*yuufeeni*
2	*sieti*	*sieti*	*yuufeeni*	*yuufeeni*
3.F	*sieti*	*seð*	*yuufeeni*	*yuufumi*
3.M	*seð*	*seð*	*yuufumi*	*yuufumi*

TABLEAU A.1 – Morphome en dhaasanac (Corbett, 2010) (données de Tosco (2001), tableau adapté de Walther (2013b))

A.1.2 Approches inférentielles : morphologies autonomes

Distinguer explicitement un niveau morphologique d'un niveau syntaxique impose de proposer un modèle de la formation des mots qui soit indépendant de la syntaxe, et qui peut donc reposer sur des mécanismes de nature différente de ceux à l'œuvre au niveau syntaxique. On est alors en droit de questionner la notion même de morphème, avatar au niveau morphologique de la compositionnalité sémantique répandue (mais pas généralisée) au niveau syntaxique. Si Hockett (1954) notait déjà que la notion de morphème relevait plus d'une commodité de description que d'une notion linguistique convaincante, Matthews (1972) montre clairement qu'elle ne résiste pas aux faits. Rejeter la notion de morphème n'implique naturellement pas obligatoirement de rejeter toute notion de structure interne des mots (Anderson, 1992 ; Stump, 2001). Mais cela conduit à reconsidérer la notion de lexique, et notamment à faire revenir le mot — en l'espèce, le mot morphologique — au premier plan. Les approches lexicales, qui consistent à modéliser directement la structure interne des mots à partir d'éléments plus petits inventoriés dans le lexique, laissent alors place à des approches qui modélisent la façon dont les mots, qui sont pourvus *a priori* de l'ensemble des traits morphologiques qu'ils expriment, sont construits en morphologie au moyen de *règles* (ou de *formules*) : ce sont les approches qualifiées d'*inférentielles* par (Stump, 2001).

L'une de ces approches est représentée par la Morphologie Articulée de Steele (1995). Elle partage avec le modèle de Lieber (1981) l'idée selon laquelle la construction d'un mot morphologique est le résultat d'un processus *incrémental*, chaque étape du processus participant à en spécifier un peu le sens, en s'appuyant à chaque étape sur le résultat des étapes précédentes : à chaque étape, le contenu informationnel est augmenté. La différence avec le modèle de Lieber (1981) réside en ceci que l'on ne part plus d'unités plus petites que le mot et inventoriées dans le lexique mais que l'on s'appuie sur des règles successives dont l'effet est de construire incrémentalement le sens du mot en parallèle à sa forme phonologique de surface : il s'agit bien d'un modèle inférentiel. Mais la Morphologie Articulée peut être critiquée au moyen d'un argument que l'on peut opposer de façon générale à toute approche incrémentale, argument qui repose sur l'abondance

d'exemples du phénomène de l'*exponence multiple*. On appelle généralement *exposant* une manifestation élémentaire (morphe, morphème, règle de réalisation…) de tout ou partie du contenu d'une structure de traits morphosyntaxiques. Or il est fréquent qu'une telle partie de structure de traits morphosyntaxiques, par exemple un trait individuel, soit exprimé par plusieurs exposants d'une forme [3]. Dans un tel cas, il n'est plus vrai que chaque étape augmente le contenu informationnel du mot en cours de construction.

Sans surprise, la dernière famille d'approches possible consiste alors à proposer des modèles qui soient à la fois inférentiels et réalisationnels. Ces modèles, tels que celui d'Anderson (1992) ou la *Paradigm Function Morphology* (PFM ; Stump, 2001, 2006) modélisent donc la construction d'un mot en morphologie comme le résultat de l'application de règles de réalisation qui, toutes, ont accès à l'ensembles des traits morphosyntaxiques à exprimer. Ces règles de réalisation sont combinées de façon séquentielle, par des compositions successives, mais cette séquentialisation est seulement liée à la façon dont le mot se construit phonologiquement et ne procède pas d'une construction itérative du sens véhiculé par le mot : le sens de la forme, et notamment la structure de traits morphosyntaxiques qu'elle exprime, est pleinement spécifiée dès le départ, et gouverne le choix des règles à appliquer. L'exponence multiple ne constitue alors plus une difficulté. Si les règles de réalisation sont purement affixales, on peut en induire une segmentation du mot en segments élémentaires. Ces segments ne sont pas des morphèmes, puisqu'ils ne sont pas porteurs d'un sens, mais sont le résultat de l'application de règles.

C'est dans ce dernier type d'approches, les approches inférentielles réalisationnelles, que se place l'ensemble de nos travaux en morphologie décrits dans ce document.

A.1.3 Approches constructives et approches abstractives

Les travaux contemporains en morphologie, par-delà les divergences entre présupposés théoriques esquissées jusqu'à présent en suivant la classification de Stump (2001), peuvent également être analysés au prisme d'une distinction plus fondamentale encore. Il s'agit d'une distinction qui relève de la façon même d'aborder les données morphologiques, et donc de l'objectif épistémologique que l'on se donne. Blevins (2006) distingue ainsi deux perspectives différentes, qu'il dénote par les termes d'*approches constructives* et d'*approches abstractives*. Pour résumer, une approche constructive vise à définir la façon dont chaque forme, individuellement, peut être dérivée à partir de briques élémentaires (morphèmes, racines, radicaux, affixes, processus, règles…). À l'inverse, une approche

3. Par exemple, dans une forme comme *iront*, le choix du radical *i-* exprime d'une part une valeur lexicale, celle d'ALLER, et le trait de futur, seules les formes futures et conditionnelles d'ALLER utilisant ce radical. Pourtant, le *-r-* qui vient ensuite exprime aussi le futur : il n'y a donc pas ici de spécification supplémentaire du sens final par ce *-r-*. Pire encore, le *-ont* exprime de façon jointe le trait de futur et le trait de troisième personne du pluriel (Matthews, 1974).

abstractive part de l'ensemble des formes et cherche à décrire les relations qu'elles entretiennent entre elles. En un sens, les approches constructives partent des briques élémentaires pour aller vers les formes alors que les approches abstractives partent des formes et, par effet de bord, peuvent faire émerger à partir des comparaisons entre formes des motifs récurrents plus élémentaires, parmi lesquels on pourrait vouloir rechercher des éléments correspondant par exemple aux notions de radicaux ou d'affixes [4]. Une approche abstractive n'est donc pas incompatible avec l'idée d'identifier dans les formes des éléments plus petits — c'est du reste ainsi que l'illustre Bonami (2014, p. 16–17) sur un exemple simple —, à condition que ces éléments soient considérés comme des abstractions que l'on a fait émerger à partir des formes et non pas des éléments constitutifs de ces formes, point de vue qui relève des approches constructives. Comme l'indique Blevins (2006), « les approches constructives font l'hypothèse que les unités élémentaires d'un système grammatical sont des segments minimaux et que le lexique des classes ouvertes consiste, au moins pour la plus grande partie, en des racines et des exposants (ou des règles qui encapsulent des exposants), mais pas en des formes complètes qui contiennent ces éléments. Une approche abstractive fait l'hypothèse que le lexique consiste avant tout en des formes complètes, desquelles sont abstraites des parties récurrentes » [5] (plus généralement des motifs récurrents). Une approche constructive s'intéresse ainsi à la façon dont les formes sont dérivées à partir de briques élémentaires là où une approche abstractive s'intéresse avant tout aux relations entre formes.

La distinction entre approches constructives et abstractives [6] est orthogonale à celles discutées jusqu'à présent : des approches comme celles de Lieber (1981), Anderson (1992) ou Stump (2001) sont constructives, tout comme la Morphologie Distribuée, alors que les approches classiques (par exemple Paul, 1880 ou (de Saussure, 1916)) ou celle de Blevins (2006), quoique de façon différente, sont abstractives.

Mais en réalité, on peut se demander si approches constructives et approches abstractives doivent ainsi être opposées. Une approche constructive a pour objet de proposer un encodage de l'inventaire des formes d'un certain nombre de lexèmes connus qui représente au mieux, quoi que cela signifie, les régularités observables. Le plus

4. Ainsi Kuryłowicz (1945) cité par Blevins (2006) : « Car la notion du thème est postérieure aux formes concrètes composant le paradigme : on trouve le thème en dégageant les éléments communs à toutes les formes casuelles du paradigme (quand il s'agit de la déclinaison). P. ê. *lup-us*, *-i*, *-o*, *-um*, *-orum*, *-is*, *-os* fondent le thème *lup-*. » Plus récemment, Bonami (2014, p. 23) explique, concernant la flexion verbale du français, que « c'est le même faisceau d'observations qui amène à conclure que l'imparfait 1PL et l'imparfait 2PL entrent dans une relation implicative systématique caractérisée par le patron d'alternance `Xjɔ̃` ⇌ `Xje`, et que la séquence `-jɔ̃` exprime l'ensemble de propriétés {imparfait, 1, PL } alors que la séquence `-je` exprime l'ensemble de propriétés {imparfait, 2, PL }. »

5. *Constructive approaches assume that the basic units of a grammatical system are segmentally minimal, and that the open-class lexicon consists, at least for the most part, of roots and exponents (or rules that encapsulate exponents), but not full word forms that contain these elements. An abstractive approach assumes that the lexicon consists in the main of full forms, from which recurrent parts are abstracted.*

6. Pour une description et une discussion plus complète de cette distinction, on pourra se référer par exemple à Bonami (2014).

souvent, cela passe par une modélisation de ce que Bonami (2014), suivant Wurzel (1984), appelle les *relations d'exponence*, c'est-à-dire « le lien qui unit un ensemble de propriétés morphosyntaxiques au matériau phonologique qui exprime ces propriétés dans les mots qui les possèdent ». Un tel encodage permet à la fois de construire l'inventaire des formes mais aussi d'extraire d'une part des propriétés de l'encodage lui-même et d'autre part des propriétés de l'inventaire de formes ainsi construit, propriétés qui rendent possible l'analyse de formes nouvelles — c'est du reste ainsi que fonctionnent les analyseurs morphologiques développés en traitement automatique des langues. À l'inverse, une approche abstractive a pour objet d'extraire, à partir d'inventaires de formes, des propriétés de cet inventaire, telles que des motifs récurrents dans ces formes, y compris les propriétés de leur répartition dans les paradigmes (Bonami et Luís, 2013). Les travaux reposant sur une approche abstractive se sont toutefois concentrés sur l'analyse des distributions de probabilité mesurant la prédictabilité entre formes, y compris entre parties récurrentes identifiées dans différentes cases : il s'agit de ce Bonami (2014) appelle les *relations implicatives*. Mais ceci ne peut se faire qu'en formalisant ce qu'est une relation entre formes, c'est-à-dire au moyen d'un encodage de ce que peuvent être les opérations de correspondance formelle nécessaires au passage d'une forme à une autre, point évoqué au chapitre 4 [7]. Autrement dit, les deux types d'approches reposent sur des présupposés arbitraires et ne se distinguent réellement que par leurs objectifs scientifiques primaires. Après tout, à partir par exemple d'un modèle (trop) simple de la morphologie qui serait purement concaténatif, les radicaux verbaux du latin que l'on pourrait définir dans une perspective constructive ne seraient pas plus arbitraires que les abstractions correspondantes que l'on pourrait obtenir avec un point de vue abstractif : il y a de bonnes chances pour que l'on aboutisse à la même chose à cet égard. Simplement, un travail de type constructif modélisant les relations d'exponence permettra de produire les paradigmes complets, au moyen d'un modèle dont la pertinence psycholinguistique n'est ni recherchée ni vérifiable, là où un travail de type abstractif étudiant les relations implicatives permettra de quantifier des relations implicatives entre formes d'une façon qui dépend complètement du modèle sur lequel on s'appuie (ici, un modèle purement concaténatif). Le modèle utilisé est ainsi une composante arbitraire commune aux deux types de points de vue. En revanche, l'arbitraire d'un travail de type constructif réside également dans les choix mis en œuvre dans une description particulière d'un système morphologique, là où un travail abstractif fait usage d'hypothèses également arbitraires qui sont faites sur ce que sont les connaissances utilisées lorsque, par exemple, on cherche

7. Ainsi, l'exemple proposé par Bonami (2014) est introduit comme suit : « Supposons établi qu'en français la variation flexionnelle intervient à la marge droite des mots. » On suppose donc une morphologie strictement concaténative, strictement affixale, et sans aucune morphophonologie. On voit bien qu'une telle analyse est loin d'être indépendante de tout présupposé théorique ou de toute hypothèse formelle.

à prédire une forme encore jamais rencontrée appartenant un lexème donné ou lorsque l'on cherche à analyser une forme nouvelle, c'est-à-dire rencontrée pour la première fois [8].

Mais comme le fait remarquer Bonami (2014, p. 24–26), les liens forts entre approches constructives et relations d'exponence, d'une part, et approches abstractives et relations implicatives, d'autre part, sont le résultat de l'histoire de la discipline et non de contraintes théoriques. Quand bien même les relations implicatives ne seraient pas mises en avant dans les approches constructives, elles peuvent toujours être extraites d'une modélisation des relations d'exponence — à condition, néanmoins, de disposer d'une notion suffisamment générale de ce que peut être une relation d'exponence. De plus, nombreuses sont les approches rendant certaines relations implicatives explicites, soit au moyen de règles de renvoi spécifiant qu'une case est remplie de la même façon qu'une autre, plutôt que de décrire comment elle est remplie, soit en modélisant explicitement des relations plus complexes entre cases ou groupes de cases (c'est le cas en Alexina$_{\text{PARSLI}}$, comme vu au chapitre 2, par exemple pour la modélisation de l'allomorphie radicale). En réalité, dans les descriptions morphologiques reposant sur les modèles et formalismes Alexina, PARSLI et Alexina$_{\text{PARSLI}}$ décrites au chapitre 2, modèles et formalismes que l'on pourrait qualifier d'orientés-constructif, l'input du processus de réalisation est presque systématiquement une forme fléchie (la forme de citation), certes complétée d'informations supplémentaires. Ainsi, l'intégralité des relations implicatives entre formes complètes n'est pas modélisée explicitement, loin de là — sauf dans le cas de paradigmes à deux cases seulement — mais le paradigme est construit au moyen de relations implicatives [9].

Plus fondamentalement, les généralisations auxquelles on peut vouloir arriver à partir des paradigmes sont tout autant accessibles, quoique différement, avec les deux types d'approches. Bonami (2014, p.23–24) indique qu'il n'est pas possible d'extraire à partir des relations implicatives des généralisations telles que le fait qu'en espagnol, les noms dont le singulier est en *-o* sont majoritairement masculins et ceux en *-a* majoritairement féminins. Mais cela n'implique pas qu'une approche abstractive ne permette pas de faire la généralisation pertinente, une approche abstractive permettant tout à fait l'induction de généralisations sur les relations d'exponence (Bonami et Luís, 2013). À l'inverse, Blevins (2006), cité par Bonami (2014), présente le cas de la première déclinaison dans la flexion

8. Par exemple, dans une approche faisant usage de *parties principales* (sous-partie du paradigme conçue pour permettre d'en inférer toutes les autres formes), le choix de l'inventaire de ces parties principales est l'un de ces arbitraires.

9. Prenons un exemple simple relevant du formalisme Alexina et de son application au français dans le lexique Le*fff*. La grammaire morphologique indique (a) que la forme de citation est la forme de l'infinitif ; (b) que cette forme se caractérise (dans le cas des verbes du premier groupe) par un suffixe *-er* ; et (c) que, par exemple, la forme IND.PRES.1PL se caractérise par un suffixe *-ons*. Les outils de flexion et d'analyse morphologique qu'Alexina permet alors de construire appliquent donc presque exactement une relation de la forme X*er* ⇌ X*ons*, que l'on peut voire comme une relation implicative. Naturellement, toutes les cases sont ici produites à partir d'une seule, et on est loin d'avoir rendu explicites toutes les relations implicatives. Mais cela aurait conduit à une redondance importante dans les descriptions.

nominale en same du nord (finno-volgaïque, ouralique). Comme le montre la table A. 2, on observe une distribution morphomique entre formes à radical géminé et formes à radical non-géminé, mais sans déterminisme quant à celle des deux sous-parties du paradigme qui est à radical géminé et celle qui est à radical non-géminé. En termes de relations implicatives, ces deux paradigmes sont identiques, en ce sens que la connaissance d'une des formes permet de savoir de façon non-ambigüe le caractère géminé ou non des radicaux utilisés dans toutes les autres cases : pour simplifier, ces deux paradigmes ressortent de la même classe flexionnelle. Bonami (2014) explique alors que la gémination n'est pas informative en termes d'exponence, puisqu'il n'y a aucune corrélation entre la présence dans une forme d'une consonne géminée et la case du paradigme à laquelle elle appartient. Mais cela n'implique pas qu'une approche constructive ne permette pas de faire la généralisation pertinente. Pour cela, il convient toutefois de se doter de mécanismes réalisationnels qui sont plus généraux que ceux généralement utilisés dans les approches réalisationnelles classiques, comme par exemple en PFM (Stump, 2001) [10]. Si, par exemple, comme en Alexina[PARSLI], on a recours à des opérations réalisationnelles dont l'effet sur l'input dépend de cet input, on peut définir une opération d'inversion de gémination qui gémine une non-géminée et remplace une géminée par la non-géminée correspondante. Une fois une telle opération définie, on peut décrire par une seule classe flexionnelle les deux paradigmes à partir de la forme de NOM.SG, en spécifiant que cette opération intervient dans la réalisation des formes de pluriel et de la forme d'accusatif singulier.

	BIHTTÁ 'morceau'		BARGU 'travail'	
	SG	PL	SG	PL
NOM	*bihttá*	*bihtát*	*bargu*	*barggut*
PFV ACC	*bihtá*	*bihtáid*	*barggu*	*bargguid*
ILL	*bihttái*	*bihtáide*	*bargui*	*bargguide*

TABLEAU A.2 – Paradigmes partiels de deux noms de la flexion du same du nord (Blevins, 2006) (données de Bartens (1989), tableau adapté de Bonami (2014))

Ainsi, la différence fondamentale entre approches abstractives et approches constructives n'est ni dans la quantité d'arbitraire qu'elles mettent en œuvre, et notamment ni dans l'importance des modèles sous-jacents, ni dans la capacité à extraire des généralisations pertinentes, dès lors que l'on généralise ce que peuvent être des règles de réalisation. Le point crucial nous semble résider dans le statut épistémologique — et donc peut-être le statut cognitif — accordé à la notion de schème flexionnel (ou de classe flexionnelle, pour

10. Bien que PFM ne contraigne pas la nature des fonctions décrivant les opérations réalisationnelles, on constate en pratique que les opérations utilisées dans les descriptions PFM sont presque toujours (quasiment) concaténatives.

simplifier), c'est-à-dire sur la nature de ce que recouvrent les propriétés systémiques du niveau flexionnel : une approche constructive suppose l'existence *a priori* d'un système flexionnel, là où une approche abstractive fait émerger *a posteriori* des propriétés systémiques (cf. section 4.3). Une fois encore, ces deux points de vue ne sont pas irréductibles l'un à l'autre : faire émerger des propriétés systémiques de façon abstractive n'est pas incompatible avec le fait de conférer à ces propriétés systémiques un statut autonome qui puisse être mis en œuvre de façon constructive pour la création de formes inconnues, par exemple dans un contexte d'apprentissage de la langue, d'intégration d'emprunts (Walther et Sagot, 2011b) ou de construction de néologismes, ou, en diachronie, pour le remplacement de formes par d'autres. Nous discutons de ces problématiques en conclusion du chapitre 4, sous l'angle de l'étude de la complexité des systèmes morphologiques, ou plus précisément des différents types de complexité que l'on peut définir et de leur relation avec l'acquisition du langage et les évolutions diachroniques.

A.1.4 Morphologie formelle, morphologie typologique et morphologie computationnelle

Du point de vue de la linguistique théorique en général et de la typologie linguistique en particulier, l'objectif principal des travaux en morphologie flexionnelle réside dans la description et la comparaison entre les systèmes flexionnels de différentes langues. Comme indiqué en introduction de cette section, un tel objectif ne peut être atteint que par des approches interdisciplinaires. En effet, pour réaliser de façon cohérente des tâches telles que la simple production automatique de paradigmes flexionnels à partir d'une description de la morphologie d'une langue, la description et la mesure des régularités et irrégularités dans ces paradigmes, ou même les études comparatives entre langues diverses, seules sont satisfaisantes des approches combinant linguistique formelle et linguistique computationnelle, appliquées à des données représentatives : la formalisation permet de garantir la cohérence d'une analyse, notamment lorsque l'on modélise le système morphologique complet d'une langue, et l'implémentation permet de vérifier concrètement la validité de l'analyse. De plus, une implémentation à grande échelle, c'est-à-dire intégrant un lexique à large couverture, permet de vérifier l'exhaustivité de l'analyse, notamment parce que cela constitue un moyen de mesurer la pertinence globale d'une description morphologique complète et l'importance relative d'un phénomène donné par rapport au système morphologique dans son ensemble.

Cette approche formelle et computationnelle reste pourtant relativement rarement mise en œuvre en morphologie théorique. Dans de nombreux cas, les formalisations sont approximatives ou ne concernent qu'un phénomène particulier, souvent indépendamment du système morphologique global duquel il est extrait (Walther, 2013b). De plus, peu de modèles disposent d'implémentations utilisables pour valider des hypothèses théoriques.

Parmi eux on peut citer PFM (Stump, 2001) et *Network Morphology* (Corbett et Fraser, 1993 ; Brown et Hippisley, 2012), associés respectivement d'une part au système *Cat's Claw* développé par Finkel [11] et d'autre part au formalisme DATR (Evans et Gazdar, 1989) et à ses extensions ultérieures telles que KATR (Finkel et Stump, 2002). Néanmoins, il existe peu d'implémentations à grande échelle dans ces systèmes. Par exemple, les analyses disponibles sur le site internet de *Cat's Claw* impliquent rarement plus d'une cinquantaine d'entrées lexicales. L'une des exceptions est l'analyse des noms du russe développée par Brown et Hippisley (2012), qui contient 1 500 entrées lexicales [12].

En morphologie computationnelle, la plupart des travaux reposent sur des approches par automates finis (Beesley et Karttunen, 2003), lesquelles n'ont aucune difficulté à construire automatiquement des paradigmes flexionnels complets et valides. En réalité, il a été montré par Karttunen (2003) que si l'on réduit les principales théories morphologiques, y compris PFM et *Network Morphology*, à leur seule capacité à produire des paradigmes flexionnels, elles peuvent être ramenées à des systèmes réalisationnels équivalents aux automates finis. Cependant, même si de telles approches computationnelles atteignent parfaitement cet objectif, elles sont souvent critiquées par les théoriciens en ce qu'elles négligent ce qu'il y a de plus intéressant et de plus important d'un point de vue théorique, à savoir la modélisation explicite des régularités et des irrégularités dans les paradigmes.

A.2 Développement de lexiques flexionnels

L'intégration de la morphologie flexionnelle au sein de dictionnaires ou de lexiques consiste à renseigner des informations permettant de spécifier explicitement ou implicitement les paradigmes complets des lemmes considérés, notamment *via* des formes caractéristiques (ou *parties principales*) ou des patrons flexionnels (par exemple par la donnée d'un lemme réputé connu et à la flexion similaire). Le développement de telles ressources est donc étroitement liée à la lexicographie et à la composition de dictionnaires, ces derniers se concentrant toutefois avant tout sur les aspects sémantiques et parfois étymologiques, en intégrant pour certains (et ce dès l'époque sumérienne tardive) des exemples d'usage.

A.2.1 Les premières ressources lexicales morphologiques

C'est à Papias, lexicographe italien du onzième siècle, que l'on attribue souvent la première entreprise de cette nature, au sein de son dictionnaire monolingue latin, l'*Elementarium Doctrinæ Rudimentum* (Boulanger, 2003) [13]. Environ 10% des articles de

11. http://www.cs.uky.edu/~raphael/linguistics/claw.html.

12. http://networkmorphology.as.uky.edu.

13. En réalité, parmi les nombreuses tablettes découvertes à Ebla (aujourd'hui Tell Mardikh, en Syrie) et documentant l'éblaïte, langue sémitique archaïque, ont été découvertes de nombreuses tablettes à contenu

ce dictionnaire, ceux pour lesquels l'auteur l'a jugé pertinent, contiennent en effet des informations telles que le genre des noms ou des informations sur la conjugaison des verbes, sous la forme d'indications permettant d'en reconstituer les parties principales (Boulanger, 2003 ; Merrilees, 1994).

De Papias au Littré et au Trésor de la Langue Française, la façon de représenter l'information flexionnelle dans les dictionnaires n'a pas significativement changé. Il s'agit même parfois d'un à-côté qui n'est pas du ressort des lexicographes qui rédigent les définitions et choisissent les exemples d'usage, comme l'indique Fradin (p.c.) à propos du Trésor de la Langue Française [14].

Pour le français, la première ressource flexionnelle conçue en tant que telle, du moins celle qui a eu le plus grand impact, est peut-être *Le Véritable Manuel des conjugaisons ou la science des conjugaisons mise à la portée de tout le monde* publié par les frères Bescherelle en 1842, ouvrage dont les rééditions puis évolutions en font encore aujourd'hui une référence. Des tableaux de conjugaison numérotés permettent, grâce à un lexique associant infinitifs et numéros de tableaux, de savoir comment fléchir les milliers de verbes qui y sont recensés.

A.2.2 Les informations lexicales morphologiques pour le TAL : des automates aux lexiques

Pour les applications de traitement automatique des langues, le développement de lexiques morphologiques n'est pas nécessairement perçu comme une priorité. Les premiers travaux en traitement automatique des langues, qui, dans les années 1950, portaient avant tout sur la traduction automatique (en priorité du russe vers l'anglais, pour des raisons évidentes), reposaient souvent sur des analyseurs morphologiques, notamment à base d'automates finis (cf. par exemple Vauquois *et al.*, 1965). Ces automates étaient associés à des ressources lexicales restreintes, couvrant notamment les exceptions, mais le développement de lexiques flexionnels à proprement parler était probablement inadapté, ne serait-ce qu'en raison de leur empreinte mémoire. Par la suite, notamment sous l'influence des travaux en phonologie formelle initiés par Chomsky et Halle (1968), l'intérêt de cette approche n'a pas été démenti, et a notamment montré sa pertinence pour des langues comme le finnois ou le turc en raison du nombre considérable de formes que l'on peut produire à partir d'un même lemme. Ces travaux ont conduit à

lexical, qui datent d'environ 2500 à 2400 avant notre ère. Il s'agit de listes monolingues sumériennes (un isolat) et éblaïtes, ainsi que de listes bilingues sumériennes-éblaïtes, plus tardives. Parmi certaines de ces tablettes, on trouve ce que l'on pourrait appeler des paradigmes morphologiques partiels (Boulanger, 2003, 87). Mais il ne s'agit pas à proprement parler de ressources lexicales morphologiques.

14. Nous n'avons pu trouver pour ainsi dire aucune information sur la façon dont ont été développées les informations flexionnelles dans le Trésor de la Langue Française (cf. par exemple Martin, 1969), informations qui ont pourtant été renseignées puisque d'une part il est possible, dans la version informatisée du dictionnaire (le TLFi), de faire des requêtes à partir de formes fléchies, et que d'autre part le lexique morphologique Morphalou, sur lequel nous reviendrons ci-dessous, en a été extrait.

l'utilisation massive d'automates finis pour la modélisation de la morphologie ainsi qu'au développement d'outils associés, destinés et à la construction et à l'analyse automatique de formes fléchies (Koskenniemi, 1984 ; Kaplan et Kay, 1994 ; Karttunen *et al.*, 1996 ; Beesley et Karttunen, 2003 ; cf. également la section A.1.4). Mais ces outils, dont en premier lieu ceux développés à XEROX, sont longtemps restés propriétaires, et il faudra attendre les années 2000 pour que des plateformes efficaces et librement disponibles pour la manipulation d'automates finis soient mises à la disposition de la communauté.

L'une des limites majeures de telles approches est qu'elles nécessitent des aménagements importants ou des encodages peu motivés linguistiquement pour traiter de systèmes morphologiques fortement non concaténatifs tels que la réduplication ou encore la morphologie gabaritique caractéristique des langues sémitiques (Lavie *et al.*, 1990). De plus, les descriptions morphologiques développées dans ce cadre sont parfois délicates à interpréter et à mettre au point. Enfin, les limitations liées à l'empreinte mémoire de lexiques fléchis (ou *extensionnels*) ont été progressivement levées, tant sur le plan algorithmique, par exemple avec l'invention du *trie* [15] (ou *arbre à lettres*, *arbre préfixe*) par De La Briandais (1959), que sur le plan du matériel lui-même.

Il faut attendre les années 1980 voire 1990 pour qu'apparaissent dans les milieux académiques des lexiques flexionnels à grande échelle. Citons, pour le français, ceux d'entre eux qui sont aujourd'hui librement disponibles [16]. L'une des premières est le DELA, développé au LADL autour de Maurice Gross à partir de plusieurs dictionnaires papier et de mots trouvés en corpus, et grâce à une description formelle et implémentée de la morphologie du français (Courtois, 1990 pour les formes simples, Silberztein, 1990 pour les mots composés). Le lexique BruLex, développé avant tout pour la psycholinguistique (Radeau *et al.*, 1990), est un prédécesseur de Lexique.org (Matos *et al.*, 2001 ; New, 2006). Le lexique de l'ABU (la Bibliothèque Universelle [17]), dont l'origine ne se laisse pas aisément déterminer, semble dater de 1999. Le lexique Morphalou, publié par l'ATILF (Romary *et al.*, 2004), est issu, via le lexique TLFnome [18] de la nomenclature Trésor de la Langue Française, moyennant une « réorganisation structurelle des données et une normalisation des étiquettes grammaticales, sans perte d'informations linguistiques ». Par la suite, outre le Le*fff*, pour lequel on pourra se reporter aux chapitres 3 et 5, on peut également citer le

15. Le terme de *trie* a été forgé par Fredkin (1960).

16. Ainsi, nous ne mentionnons pas dans le corps du texte le lexique BDLex (de Calmès et Pérennou, 1998), lexique payant qui inclut également des informations phonétiques pour permettre des applications liées au traitement automatique de textes mais aussi de la parole. On peut noter que ce lexique fait l'objet d'une publication dès 1987 dans les actes de la *European Conference on Speech Technology*, publication qui n'est accessible en ligne, sur le site des actes, qu'au moyen d'un mot de passe. Ne s'agissant pas d'un article de journal mais d'un article de conférence, l'impossibilité d'un accès direct et gratuit à cet article nous amène, par principe, à l'ignorer.

17. `http://abu.cnam.fr`

18. Développé par Jacques Maucourt et Marc Papin en 1996 (cf. Bernard *et al.*, 2002).

lexique GLÀFF (Gros Lexique À tout Faire du Français), extrait du Wiktionnaire [19] (Sajous *et al.*, 2013 ; Hathout *et al.*, 2014 ; cf. également la section 3.1.3).

S'agissant de langues autres que le français, on peut citer notamment CELEX (Burnage, 1990 ; Baayen *et al.*, 1993), qui est un lexique couvrant l'anglais, l'allemand et le néerlandais, sans toutefois que les données lexicales pour ces trois langues soient reliées entre elles. Ce lexique inclut des informations flexionnelles, mais couvre également d'autres niveaux d'analyse, notamment les niveaux phonétiques, et syntaxiques. Mais cette base n'est pas librement disponible.

On peut remarquer à ce stade que le traitement automatique des langues, s'il bénéficie crucialement des informations fournies par un lexique morphologique, n'a pas nécessairement besoin que ce dernier soit le plus volumineux possible. Illustrons ceci sur le français. La présence de nombreux mots rares, anciens ou techniques, typique de lexiques comme Morphalou, peut induire des ambiguïtés inutiles (le verbe ANSER 'garnir d'une anse' a ainsi plusieurs formes fléchies homographes avec celles du nom ANSE). Dans un système de correction ou de normalisation orthographique, de tels mots peuvent également induire en erreur en constituant des résultats possibles quoique peu vraisemblables [20]. C'est d'autant plus vrai si, comme dans le Wiktionnaire et donc dans le GLÀFF, sont ajoutées des néologismes quasiment non attestés (ainsi le verbe AFTERSHAVER 'mettre de l'aftershave', dont plusieurs formes fléchies sont homographes avec le nom AFTERSHAVE) [21]. À l'inverse, on voit mal, mis à part alourdir les traitements, l'intérêt de disposer dans un lexique destiné au traitement du français contemporain d'orthographes archaïques (ADHÆRER pour ADHÉRER est dans le GLÀFF), voire d'entrées rarissimes (souvent techniques et/ou dialectales) dont la forme permettrait une analyse dynamique si d'aventure on la rencontrait en corpus (cf. section 3.2). La quête irraisonnée de l'exhaustivité dans les lexiques morphologiques n'est donc pas nécessairement une bonne chose.

Un point commun important entre un système faisant usage d'automates finis et un système faisant usage d'un lexique flexionnel est la nécessité de disposer d'une grammaire morphologique : une telle grammaire peut être utilisée pour fléchir statiquement un lexique morphologique ou être encodée d'une façon qui la rende

19. `http://fr.wiktionary.com`.

20. Baranes (2015) donne l'exemple de « l'archaïsme *affin* qui, statistiquement, a plus de chance [de devoir être corrigé en] *afin* ou *affine* [que d'être correct]. […]. Si [cet archaïsme] est dans notre lexique et qu'on rencontre la forme erronée *afiin*, un système hésitera entre les formes *afin* et *affin* pour la corriger. »

21. Baranes (2015) rappelle pourtant qu'un article est considéré en principe comme pertinent par les administrateurs du Wiktionnaire à condition qu'il respecte les conditions suivantes :

- il décrit des termes de la langue réellement utilisés (anciennement ou actuellement),
- ces termes peuvent être attestés avec des sources sérieuses (limite basse),
- ils ne sont pas de simples utilisations de termes plus simples qui se suffisent à eux mêmes (limite haute).

utilisable dynamiquement (par exemple sous la forme d'un automate) afin d'analyser toutes les formes à traiter ou, dans le cas où l'on utilise un lexique, seulement les formes qui ne sont pas connues de ce dernier. Les langues dont il sera question dans ce document permettent toutes la construction statique d'un lexique morphologique extensionnel, y compris lorsque l'on dispose d'un lexique (intensionnel) de grande taille (nous n'avons pas travaillé sur des langues telles que le turc ou le finnois). Comme discuté à la section 3.2, ceci n'est nullement incompatible, lors de l'analyse d'un corpus, avec une analyse automatique des formes qui sont inconnues du lexique.

Les ressources et outils traitant de la morphologie flexionnelle reposent donc principalement sur deux types de connaissances linguistiques, comme discuté au chapitre 2 : un inventaire de formes de citation associées à des propriétés flexionnelles et/ou une grammaire morphologique. Nous l'avons vu, la première approche pour le développement de telles ressources a été une approche lexicographique et manuelle. Ces ressources, lorsqu'elles existent déjà et sont exploitables informatiquement, peuvent servir de point de départ au développement de lexiques morphologiques à grande échelle, souvent par l'intermédiaire de procédés permettant d'extraire des informations structurées et formalisées à partir de ressources qui le sont moins. Mais l'acquisition automatique ou semi-automatique d'informations lexicales morphologiques peut constituer une alternative utile, en particulier pour pallier l'incomplétude des ressources existantes, notamment pour la prise en compte d'entrées lexicales récentes ou spécialisées, ainsi que dans le cas de langues peu dotées. Naturellement, la qualité des ressources ainsi construites est variable. Elle dépend du volume de travail manuel que l'on met en œuvre au cours du processus, mais avant tout de la nature des informations dont l'on dispose au départ. On peut noter à cet égard que, selon le contexte d'utilisation de telles ressources flexionnelles, leur qualité a un impact plus ou moins important. La présence de formes erronnées ou manquantes, pour peu qu'elles soient rares, aura ainsi un effet moins important pour des tâches d'extraction d'informations que pour des tâches de génération de textes. Nous discutons aux chapitres 8 à 7 de la façon dont de telles ressources peuvent améliorer des résultats en étiquetage morphosyntaxique ou en analyse syntaxique.

A.3 Développement de lexiques dérivationnels

Il existe désormais de nombreuses ressources lexicales à grande échelle comportant des descriptions morphologiques flexionnelles précises pour de nombreuses langues, développées manuellement ou, le plus souvent, avec l'aide de techniques automatiques telles que celles décrites précédemment. Mais il est plus rare que soit également couverte la morphologie constructionnelle. Cette dernière est pourtant la partie de la morphologie

qui permet de créer des lexèmes à partir d'autres lexèmes, à la fois dans la langue générale et dans les langues de spécialité, et d'augmenter ainsi le lexique d'une langue. Il s'agit donc d'un procédé de structuration du lexique qui relie des lexèmes entre eux et un procédé d'extension du lexique. À ce double titre, la modélisation des opérations dérivationnelles ou la disponibilité de ressources lexicales fournissant des informations de dérivation est cruciale à la fois pour la description linguistique et pour le traitement automatique des langues (TAL). En linguistique, les ressources lexicales enrichies de liens dérivationnels permettent une approche systémique du lexique et des procédés productifs de construction de lexèmes. En traitement automatique des langues, de nombreux travaux antérieurs ont ainsi montré l'importance de la prise en compte de la morphologie constructionnelle pour l'analyse des mots inconnus (néologismes, termes techniques) et l'enrichissement de ressources lexicales (Dal *et al.*, 1999 ; Hathout et Tanguy, 2005 ; Sagot *et al.*, 2013, cf. section 3.2), l'analyse syntaxique (Bourigault et Frérot, 2004), mais également dans des contextes plus applicatifs tels que les systèmes de question-réponse (Bernhard *et al.*, 2011) ou de traduction automatique (Cartoni, 2009). Par ailleurs, la morphologie dérivationnelle peut également contribuer à l'enrichissement de ressources lexicales au niveau syntaxique (par exemple, au moyen d'informations de sous-catégorisation [22]) et au niveau sémantique (par exemple, pour l'extension de ressources de type wordnet, cf. Sagot *et al.*, 2009a, 2008, 2009b).

A.4 Quantifier et mesurer la complexité morphologique

Comme indiqué au début du chapitre 4, quantifier et mesurer la complexité morphologique est devenu un enjeu important des travaux en morphologie. À cet égard, différentes approches ont été proposées dans la littérature, que l'on peut classer en trois grandes familles : d'une part les approches par comptage et d'autre part les approches reposant sur la théorie de l'information, parmi lesquelles les approches entropiques et les approches reposant sur des descriptions morphologiques.

A.4.1 Approches par comptage

Les plus simples de ces approches, les *approches par comptage* (McWhorter, 2001 ; Bickel et Nichols, 2005 ; Shosted, 2006), se contentent de compter les occurrences d'un ensemble de propriétés linguistiques définies manuellement, et notamment la taille de divers inventaires : inventaires des phonèmes, des catégories, des cas, des types de morph(ème)es, des classes flexionnelles, etc.

22. On pourrait ainsi tenter d'induire des cadres de sous-catégorisation pour des noms prédicatifs dérivés de verbes dont on connaîtrait les propriétés syntaxiques.

Langue	français	russe	hongrois	swahili
Nombre de genres	2	3	aucun	$\geq$5
Nombre de cas	aucun	6–7	>10	aucun

Tableau A.3 – Nombre de genres et de cas dans quatre langues (données de WALS). On se convainc facilement, au vu de ces données, que le russe a une morphologie plus « complexe » que le français. En revanche, comparer le hongrois au swahili est impossible, tout comme l'est une comparaison de ces langues au russe.

De telles approches, qui cherchent à comparer plusieurs langues entre elles quant à la complexité de leurs systèmes morphologiques, sont intrinsèquement discutables (voir par exemple Blevins, 2006 ou Sagot, 2013a) : tant l'ensemble de propriétés choisies que les critères selon lesquels ces propriétés sont décrites sont très difficiles à définir selon des principes clairs, objectifs et reproductibles, et chaque typologue ou morphologue pourra s'appuyer sur des choix différents. Ainsi, comment construire d'une façon objective et indépendante de la langue un inventaire de parties du discours ? Comment disposer d'un moyen objectif de compter le nombre de classes flexionnelles quelle que soit la langue ? Combien y a-t-il de cas dans la morphologie nominale du russe ou du slovaque ? La notion de cas est-elle comparable dans toutes les langues que l'on souhaite comparer entre elles ? Peut-on comparer le nombre de genres en français ou en allemand et le nombre de classes nominales dans les langues bantoues ? Si une langue est complexe sur un certain plan (nombre de cas élevés, par exemple) et une autre sur un autre plan (nombre de classes nominales élevés, par exemple), comment les comparer ? Un cas de plus vaut-il plus ou moins qu'une classe flexionnelle supplémentaire ? Le tableau A.3 illustre l'absurdité de telles questions. Plus fondamentalement, une telle approche part du principe que, quel que soit l'inventaire dont l'on mesure la taille, plus cet inventaire est grand, plus la complexité du système morphologique est importante. Ce qui n'est pas nécessairement justifié : on peut imaginer des systèmes flexionnels avec peu de cases mais de nombreuses irrégularités [23] et d'autres avec de nombreuses cases mais des paradigmes très réguliers.

A.4.2 Approches reposant sur la théorie de l'information

Des approches alternatives de mesure de la complexité existent, dont la plupart reposent sur des définitions de la complexité qui sont issues de la Théorie de l'Information. Dans ce cadre, deux définitions différentes, mais qui ne sont pas sans liens, ont été utilisées dans les travaux récents. Elles s'appliquent toutes deux à n'importe quel type de message et pas seulement aux descriptions ou aux modèles linguistiques : (i) l'*entropie*

23. Ainsi par exemple le système verbal du créole mauricien, qui n'a que deux cases mais qui présente un niveau élevé d'imprédictibilité (Bonami *et al.*, 2011).

informationnelle (ou *complexité de Shannon*), dont le principal inconvénient est qu'il nécessite l'encodage du message sous la forme d'une séquence de variables aléatoires indépendantes et identiquement distribuées selon un certain modèle probabiliste, ce qui est difficile en pratique (Shannon, 1948), et (ii) l'*entropie algorithmique*, ou *complexité de Kolmogorov* (Solomonoff, 1960 ; Kolmogorov, 1963 ; Solomonoff, 1964 ; Kolmogorov, 1965), un moyen plus générique et plus objectif de mesure de la quantité d'informations contenue dans un message, mais qui n'est pas directement calculable et pour lequel on doit donc se contenter d'approximations. Ces deux notions recouvrent des points de vue distincts, quoiqu'évidemment liés, sur la notion générale de complexité : pour résumer, l'entropie informationnelle s'intéresse au degré de prédictibilité d'une nouvelle instance d'un ensemble de données à étudier, alors que la complexité de Kolmogorov mesure le contenu informationnel d'un ensemble de données défini [24]. Nous allons passer rapidement en revue chacune de ces deux approches.

A.4.2.1 Approches reposant sur des descriptions morphologiques

La complexité de Kolmogorov a pour objectif de mesurer le degré d'aléatoire dans les données. Elle repose sur l'intuition suivante : une description des données capte moins bien les redondances, et sera donc considérée plus complexe qu'une autre, si elle nécessite un message plus long pour être décrite (on peut concevoir le contenu de ce message comme un ensemble de règles spécifiant comment produire les données initiales) [25]. La complexité d'un jeu de données est alors définie comme égale à celle du plus petit message permettant de les décrire. Ainsi, la complexité de Kolmogorov a l'avantage d'être plus générale et de ne pas nécessiter de modèle probabiliste sous-jacent pour les données à étudier.

Il s'agit donc d'une approche fondamentalement constructive, au sens où nous l'avons discuté au chapitre 2. Cette mesure capte ainsi les redondances structurelles au sein d'une description, qu'elles soient intuitives ou non. Cependant, le calcul de cette complexité, c'est-à-dire de la quantité d'information contenue dans le message, est impossible directement. On a donc recours à des approximations.

24. On peut illustrer la différence entre ces deux points de vue sur la complexité par un exemple simple : considérons une source émettant aléatoirement et de façon équiprobable l'un ou l'autre de deux messages pré-établis, chacun de ces messages étant très complexes. Chacun de ces messages peut ainsi avoir une complexité de Kolmogorov très élevée. En revanche, la complexité de Shannon du système est égale à 1, chacun des deux messages étant émis de façon équiprobable. Si en revanche on dispose d'une source émettant successivement des messages atomiques selon une distribution « raisonnable » et que l'on calcule non pas la complexité de Kolmogorov de chaque message mais celle de la séquence ainsi produite, alors l'entropie de Shannon est un majorant de la complexité de Kolmogorov, et tend vers cette dernière lorsque la taille de la séquence tend vers l'infini.

25. La formulation habituelle ne parle pas de *message décrivant des données* mais de *programme (destiné à une machine de Turing) produisant des donnés*. Nous présentons ici toutefois la complexité de Kolmogorov d'une manière plus directement liée à la façon dont nous l'utiliserons par la suite.

Une première façon de faire consiste à procéder à une compression sans pertes des données : la longueur du résultat de l'opération de compression est une bonne approximation de leur complexité de Kolmogorov. Ainsi, le programme de compression usuel *gzip*, qui utilise une variante de l'algorithme LZ77 (Ziv et Lempel, 1977), a été utilisé par Juola (1998) pour approximer la quantité d'informations portée par le niveau morphologique dans des corpus en différentes langues. Pour cela, il calcule la différence entre la longueur du corpus original compressé et celle de ce même corpus, mais compressé après avoir remplacé chaque mot par un nombre, de façon à priver les formes des relations formelles, et notamment morphologiques, qu'elles entretiennent entre elles. Dans nos travaux, et contrairement à Juola (1998), nous nous sommes concentrés sur l'estimation de la complexité du système morphologique en soi (description morphologique, paradigmes de formes fléchies…). Mais l'approximation de la complexité de Kolmogorov par des algorithmes de compression est une approche générique [26].

L'autre approximation souvent utilisée consiste à utiliser la notion de *longueur de description*. On part pour cela d'un un sous-ensemble particulier de descriptions possibles que l'on se donne à l'avance et que l'on nomme le *modèle* [27]. On définit alors sur cet ensemble de descriptions une *fonction de codage*, c'est-à-dire une fonction bijective permettant de transformer chacune des descriptions permises par le modèle choisi en un *code*. On fait en sorte que cette fonction capte le mieux possible les généralisations envisageables au sein de ces descriptions, et on fait l'hypothèse approximative que les codes produits à partir de descriptions valides sont des suites de symboles indépendants les uns des autres. On peut alors approcher la quantité d'information portée par une description donnée au sein du modèle choisi par la quantité d'information du code correspondant, calculée directement comme le produit de son entropie par sa longueur. Le résultat est ce que l'on appelle la *longueur de description* [28].

Naturellement, plus le modèle utilisé et la fonction de codage retenue permettent de capter des généralisations pertinentes, c'est-à-dire des structures pertinentes dans les données, plus la longueur de description sera proche de l'intuition que l'on peut avoir

26. Elle n'est toutefois générique que dans la mesure où l'on se contente des types de motifs récurrents exploités par l'algorithme LZ77 et que l'on accepte d'ignorer les autres types de régularités, dont singulièrement les régularités impliquant des sous-séquences discontinues.

27. Classiquement, ce modèle est l'ensemble des programmes valides dans un langage de programmation donné, pour peu qu'il ait la puissance d'expression d'une machine de Turing. Dans un tel contexte, le choix du langage n'affecte pas, à une constante près, la valeur de la complexité de Kolmogorov. Mais dans notre cas, une machine de Turing est un dispositif bien trop expressif pour ce que nous avons à faire. C'est ce qui justifie le fait que nos mesures de complexité dépendent, quant à elles, du modèle choisi.

28. Cette notion, qui dépend donc du modèle choisi et de la fonction de codage, est le fondement du paradigme d'apprentissage automatique appelé *longueur de description minimale* (*Minimum Description Length*, ou *MDL*) (Rissanen, 1984) : dans ce paradigme, l'apprentissage automatique se fait *via* l'identification de la description que l'on peut faire des données d'apprentissage qui ait une longueur de description minimale par rapport à un modèle choisi.

de la notion de complexité. C'est du reste par la restriction des descriptions admissibles, et donc par la façon dont l'on approxime la complexité de Kolmogorov, que l'on peut rapprocher l'intuition qui est au fondement de la complexité de Kolmogorov d'une notion intuitive de la complexité [29].

Dans notre cas, les données à traiter sont un lexique morphologique extensionnel, et la description des données dont l'on cherchera à mesurer la longueur de description sera une description formalisée d'un système morphologique permettant de génerer ce lexique extensionnel, description constituée d'une grammaire morphologique et d'un lexique morphologique associé (cf. chapitre 2). C'est cette direction de recherche que nous avons qualifiée plus haut d'*approches reposant sur des descriptions morphologiques*, et que nous avons développée et mise en œuvre sur différentes langues (cf. section 4.1). Elle conduit donc à des mesures de la complexité morphologique qui sont dépendantes du modèle et s'appliquent à des descriptions particulières, ce qui ouvre la voie à des comparaisons entre descriptions concurrentes d'un même système morphologique. Il s'agit donc d'un objectif distinct de ceux des travaux constrastifs entre langues visant à comparer leur complexité morphologique (ou linguistique) de façon globale (McWhorter, 2001 ; Juola, 2008 ; Bane, 2008). On peut remarquer enfin qu'une description morphologique est elle-même un objet structuré, ce qui n'est pas toujours exploité par les travaux antérieurs dans ce domaine [30]. Nous proposerons donc à la section 4.1 une nouvelle mesure de la complexité morphologique associée à Alexina, qui constitue notre modèle, en tant qu'il définit notre ensemble des descriptions possibles.

Dans une telle approche, la mesure de complexité est donc une mesure de *compacité* d'une description donnée, certes restreinte à un ensemble de descriptions possibles et, plus fondamentalement, à un modèle formel donné de la morphologie flexionnelle. Mais une telle mesure permet de maîtriser l'arbitraire sous-jacent aux descriptions constructives (cf. section A.1.3), et singulièrement l'arbitraire des décisions de segmentation des formes en morphes. S'appuyer sur un modèle formel riche tel qu'Alexina permet d'exprimer

29. Ce faisant, on s'écarte pourtant de la complexité de Kolmogorov à proprement parler, avec pour inconvénient d'introduire un arbitraire et pour avantage de mieux correspondre à la notion intuitive de complexité. Mais cet avantage est important. Considérons par exemple les deux séquences de chiffres suivantes : (i) 79323846264338327950288419716939937510582097494459230781640628620899862803482534211706798 2 et (ii) 39044668212534993176949622575370703990717521385899755725140494919206137510425498012326900 4. La seconde sequence est aléatoire, mais la première est constituée de 90 décimales successives de π à partir de la treizième. Il est donc possible de la produire avec un programme court, ce qui n'est pas le cas de la seconde. Autrement dit, sans modèle restreignant l'espace des descriptions possibles, ces deux descriptions sont de complexités très différentes, ce qui contredit la façon dont on perçoit intuitivement leur complexité. Et c'est bien une notion intuitive de la complexité que l'on cherche à formaliser dans le cas de l'étude de la complexité des systèmes morphologiques — quand bien même l'intuition du locuteur et celle du linguiste pourraient bien différer.

30. En particulier, l'évaluation de la complexité d'une représentation du lexique morphologique ne peut se réduire à mesurer la complexité d'un corpus dont les formes ont été segmentées en morph(ème)es. Cette approche est toutefois le fondement de travaux pionniers en acquisition automatique d'informations morphologiques (Goldsmith, 2001).

des généralisations complexes grâce au développement manuel des descriptions. À l'inverse, les approches où les généralisations sont extraites automatiquement des données ne permettant pas, ou bien plus difficilement, de capturer toutes les généralisations pertinentes. Néanmoins, le fait que les approches reposant sur des descriptions morphologiques aient pour prérequis, à ce jour du moins, le développement manuel de descriptions morphologiques en fait une approche relativement coûteuse. De plus, ces descriptions courent le risque d'être biaisées par des préconceptions sur ce que peuvent être des façons raisonnables de représenter le système flexionnel considéré. Ces inconvénients, qui sont la conséquence du point de vue constructif inhérent à ces approches, justifient néanmoins le développement d'approches fondées sur la seule analyse des paradigmes flexionnels complets.

A.4.2.2 Approches reposant sur l'entropie informationnelle au sein des paradigmes

Une autre direction de recherche, définissant une autre forme de complexité morphologique reposant sur l'entropie informationnelle (ou complexité de Shannon), a de fait été développée. On peut en effet considérer un ensemble de données (un corpus, un lexique) comme un ensemble d'instances (des phrases ou mots, des unités lexicales…) émises successivement par une source. L'entropie de Shannon est alors un moyen d'estimer le degré d'incertitude que l'on rencontre si l'on veut prédire ce que sera une nouvelle instance émise par la source en prenant en compte un modèle des données et les instances précédemment émises. Autrement dit, la mesure de complexité est ici une mesure de *prédictibilité*, c'est-à-dire de *régularité* des données. Le calcul de l'entropie de Shannon requiert de modéliser les données sous la forme d'une séquence de variables aléatoires indépendantes et identiquement distribuées qui suivent un certain modèle probabiliste, ce qui est souvent délicat voire impossible en pratique. De plus, la façon dont ce modèle probabiliste est constitué a un impact important.

Cette approche a été néanmoins utilisée directement sur des corpus (Moscoso del Prado Martín *et al.*, 2004 ; Moscoso del Prado Martín, 2011 ; Pellegrino *et al.*, 2007, 2011), mais également pour mesurer l'interprédictibilité entre cases d'un paradigme : c'est ce qu'Ackerman et ses collègues ont appelé le Problème du Remplissage des Cases d'un Paradigme (*Paradigm Cell Filling Problem, ci-après PCFP* ; Ackerman *et al.*, 2009 ; Malouf et Ackerman, 2010 ; Ackerman et Malouf, 2013 ; cf. également la section 4.2.1). Ils définissent alors la complexité d'un système morphologique comme étant la moyenne des entropies conditionnelles de chacune des cases d'un paradigme à partir de chacune des autres cases.

Cette définition de la complexité morphologique estime donc combien fiables sont les patrons implicatifs pour deviner comment remplir une case à partir d'une autre. Il s'agit donc, au moins dans les travaux cités ci-dessus et étudiant la prédictibilité entre cases, d'une appoche typiquement abstractive, au sens vu au chapitre 2. Mais ce type d'approche

n'est pas exempt de difficultés, que nous étudions à la section 4.2. Ces difficultés sont liées à la façon dont les paradigmes et les relations entre les formes qui les composent sont appréhendées et représentées, notamment si, comme le font Ackerman *et al.* (2009), on part d'un inventaire de classes flexionnelles et d'une segmentation des formes donnés *a priori*. Cela relève ainsi d'une maîtrise insuffisante de l'arbitraire descriptif : les choix ayant conduit aux décisions de segmentation et à l'inventaire de classes flexionnelles retenu affectent la pertinence des résultats obtenus. Cette part d'arbitraire ne peut être maîtrisé que par l'utilisation d'approches non-supervisées pour présider à la formalisation et à l'organisation des relations implicatives entre cases. Deux questions au moins restent donc en suspens à ce stade : comment extraire automatiquement des patrons implicatifs captant un maximum de généralistions ? et comment extraire automatiquement un inventaire, ou mieux, une hiérarchie de classes flexionnelles ? Répondre à ces questions nécessite naturellement de prendre parti sur la façon dont on peut caractériser les relations entre formes (y compris sur la place de la morphophonologie) ainsi que la notion de classe flexionnelle, deux problématiques intimement reliées.

Il ressort de cette discussion introductive que la complexité morphologique peut être définie et calculée de multiples façons, remettant ainsi en cause l'idée même de mesurer *la* complexité morphologique (voir la conclusion du chapitre 4).

A.5 Modélisation de l'information lexico-syntaxique

La représentation de la structure syntaxique des énoncés est au cœur de la description linguistique. Au sein des approches linguistiques modernes, ce souci prend schématiquement la forme d'une opposition ou d'une complémentarité entre deux types de représentations (cf. par exemple Kahane, 2001) :

1. Les représentations en dépendances, qui sont centrées sur des relations asymétriques entre mots, qui relient un mot dit régi au mot dont il dépend, auquel il se rattache. La première théorie linguistique qui s'appuie sur ce type de représentations est celle de Tesnière (1934, 1959). Les travaux de Tesnière font suite à des travaux antérieurs comme ceux de Jespersen (1924), mais on peut faire remonter certaines idées sous-jacentes au modèle en dépendances au premier grammairien connu, Pāṇini, et à sa description étonamment moderne du sanskrit (l'Aṣṭādhyāyī, $\sim$IV^{e} siècle av. J. C.). Puisqu'une représentation en dépendances met l'accent sur les relations entre les mots d'un énoncé plutôt que sur leur organisation en chaîne linéaire, il n'est pas surprenant que ce type de représentations ait été particulièrement étudié par des linguistes travaillant sur des langues où l'ordre des mots dans un énoncé jouit d'une certaine liberté. C'est par exemple le cas du tchèque (Functional Generative Description, FGD, Sgall *et al.*, 1986), du russe (Théorie Sens-Texte,

TST, Meľčuk, 1974, 1988) ou de l'allemand (Engel, 2009, 3e éd. d'un ouvrage paru initialement en 1988).

2. Les représentations en constituants, ou syntagmatiques, qui sont centrées sur la structure linéaire des mots dans la phrase. On fait traditionnellement remonter ces travaux à Bloomfield (1933a) et au distributionnalisme. C'est la linguistique générative initiée notamment par Chomsky (1957, 1965) qui assurera pendant plusieurs décennies la prééminence des modèles syntagmatiques, dont la pertinence pour des langues configurationnelles (à ordre des mots plus strictement contraint) comme l'anglais est réelle.

Toutefois, depuis le début des années 1980, les modèles génératifs de la syntaxe introduisent, sous différentes formes, des relations de type dépendancielles. C'est notamment le cas des modèles chomskiens les plus récents, depuis le Gouverment et Liage jusqu'au Programme Minimaliste, mais également de modèles génératifs partiellement lexicalisés, comme les Grammaires Lexicales Fonctionnelles (*Lexical Functional Grammars*, LFG) (Bresnan, 1982 ; Kaplan et Bresnan, 1982) ou les modèles des Grammaires Syntagmatiques que sont notamment GPSG (*Generalized Phrase Structure Grammars*, Gazdar, 1985) et HPSG (*Head-driven Phrase Structure Grammars*, Pollard et Sag, 1987). Si, en LFG, la notion de dépendance peut être identifiée dans la façon dont les propriétés de valence sont représentées et mises en œuvre par dessus un « squelette » strictement syntagmatique, un modèle comme HPSG fait le choix opposé, en modélisant avant tout les relations entre les mots et en faisant usage de mécanismes de réalisation non nécessairement triviaux, permettant ainsi de rendre compte de phénomènes rencontrés dans des langues à ordre plus libre (Reape, 1993 ; Kathol, 1995). On peut toutefois montrer que les approches suivies par LFG d'une part et HPSG d'autre part partagent fondamentalement de nombreux points communs, dont cette idée consistant à découpler ordre des mots et relations entre mots (Manning, 1995). En parallèle, les modèles strictement lexicalisés, comme les Grammaires d'Arbres Adjoints (*Tree Adjoining Grammars*, TAG, Joshi *et al.*, 1975 ; Joshi, 1987 ou les Grammaires Catégorielles (Moortgat, 1988) ont montré qu'il était possible de construire des systèmes formels permettant la construction de structures syntagmatiques selon un processus de dérivation que l'on peut représenter lui-même au moyen de structures qui, sous certaines hypothèses, s'apparentent à des structures en dépendances.

Cette évolution est la conséquence d'une double prise de conscience :

— Le lexique, dont l'importance était fortement sous-estimée par les courants majoritaires de l'époque en linguistique générative, a été remis au centre des modèles syntaxiques. On peut noter à cet égard que les travaux de l'école du Lexique-Grammaire, autour de Maurice Gross, ont précédé en France de plusieurs années cette évolution majeure de la linguistique américaine (Gross, 1975 ; Boons

et al., 1976a,b ; Guillet et Leclère, 1992), conférant à l'étude du lexique et aux spécificités idiosyncrasiques de chacune des entrées qui le compose une importance primordiale. Le postulat sous-jacent au développement du Lexique-Grammaire, pour le français d'abord puis pour d'autres langues, était précisément de montrer qu'il n'y avait pour ainsi dire aucune généralisation possible que l'on puisse identifier dans le lexique, pris ici comme inventaire des propriétés syntaxiques des unités prédicatives, et que la description exhaustive des entrées lexicales d'une langue et de ses propriétés syntaxiques était la tâche principale de tout travail de description de la langue en général. Si cette quête de l'exhaustivité et le quasi-refus de toute généralisation semblent aujourd'hui excessifs, il n'en reste pas moins qu'il s'agissait là d'une entreprise éminemment lexicaliste, préfigurant ainsi les futurs modèles tels que LFG, TAG, GPSG ou HPSG. Or une approche lexicaliste, par définition, est centrée sur l'unité lexicale et ses propriétés de combinaison avec d'autres unités lexicales, autrement dit sur des propriétés dépendancielles, telles que la valence [31].

— Les grammaires purement syntagmatiques ne rendent pas compte de façon satisfaisante des langues pour lesquelles l'ordre des mots est relativement libre, c'est-à-dire en réalité de la majorité des langues. La prééminence des études génératives sur l'anglais, langue typologiquement inhabituelle à cet égard, a influencé à l'excès les travaux en linguistique générative [32].

Dans ce contexte, le développement de *lexiques syntaxiques*, c'est-à-dire l'encodage des propriétés combinatoires des unités lexicales syntaxiques, autrement dit des dépendances qu'elle peut ou doit susciter dans un énoncé, devient une entreprise indispensable. C'est le sens même de la démarche de description linguistique de travaux tels que ceux du Lexique-Grammaire. Mais c'est également la contrepartie naturelle et incontournable de tout modèle linguistique reposant sur les dépendances, d'où par exemple les travaux lexicographiques conséquents entrepris depuis plusieurs dizaines d'années dans le cadre de la Théorie Sens-Texte (TST, en anglais *Meaning-Text Theory* ou MTT) autour du Dictionnaire Explicatif et Combinatoire (Meľčuk et Polguère, 1995 ; Meľčuk *et al.*, 1984, 1988, 1992, 1999). C'est également une nécessité pour toute implémentation d'analyseurs syntaxiques reposant sur des modèles tels que LFG, HPSG ou TAG.

31. Cette évolution, en réalité, n'est pas restreinte à la syntaxe. Bien au contraire, le lexicalisme est une approche qui remet le lexique dans son ensemble au cœur du modèle, y compris voir surtout la sémantique lexicale.

32. Aujourd'hui encore, en traitement automatique des langues, la prééminence de l'anglais sur les travaux de recherche, y compris hors des pays anglophones, est massive. Il est à cet égard intéressant de constater qu'une communauté a pu se fédérer autour de l'analyse syntaxique des langues dites « à morphologie riche » (Seddah *et al.*, 2013b), terme à comprendre comme désignant toute langue qui, en comparaison avec l'anglais (et au mandarin), a une morphologie flexionnelle non triviale et un ordre des mots moins figé. Le français, dans ce cadre, est ainsi considéré comme étant une langue à morphologie riche, ce qui est typologiquement exagéré.

Naturellement, il existe de nombreuses façons de coder l'information lexico-syntaxique, et notamment de modéliser le fait qu'un même lexème (cf. chapitre 1) permet en général la construction d'énoncés dans lesquels il a une valence différente, pour de multiples raisons (arguments non réalisés, diathèses marquées, alternances régulières entre plusieurs constructions, mais aussi comportements différents quoique régulièrement prédictibles de différentes formes fléchies du lexème, etc.). On parle alors de cadres de sous-catégorisation différents, ou plus simplement de cadres différents. Plusieurs approches sont alors possibles pour encoder l'information lexico-syntaxique. Parmi les travaux réalisés aux États-Unis, on peut en identifier trois grandes familles :

— les approches extensionnelles, où l'on spécifie explicitement tous les cadres possibles (Bresnan, 1982) ;

— les approches par règles, qui, tout en étant extérieures au lexique, permettent d'établir des liens de correspondance entre arguments sémantiques (actants) et arguments syntaxiques (arguments à proprement parler), ces derniers pouvant prendre part à divers types de cadres ; on ne peut alors pas éviter d'aborder la question de l'interface syntaxe-sémantique, au moins au niveau du lexique (Gross, 1975 ; Jackendoff, 1990 ; Bresnan, 2001) ;

— les approches constructionnelles (Borer, 2005 ; Goldberg, 1995), qui font jouer aux informations lexicales un rôle moins important, et qui modélisent l'apparition de la structure argumentale par l'intégration des unités lexicales à des constructions syntaxiques.

C'est dans la deuxième famille d'approches que se placent les travaux de Levin (1993) sur les verbes anglais, ainsi que le projet VerbNet (Kipper *et al.*, 2000), dont ils sont l'un des fondements. L'idée sous-jacente est que les cadres de sous-catégorisation d'un verbe reflètent sa sémantique. Chaque classe comporte un ensemble de verbes qui partagent les mêmes *alternations*, c'est-à-dire les mêmes ensembles de cadres, les mêmes rôles thématiques et les mêmes restrictions de sélection. Plus proche encore de la sémantique lexicale se situe un projet comme PropBank (Kingsbury et Palmer, 2002). Parce que notre objet ici est la syntaxe lexicale, nous n'insisterons pas sur ces travaux, pas plus que nous ne décrirons ici en détail FrameNet (Baker *et al.*, 1998), dont le niveau d'analyse est proprement sémantique, sans lien, en principe du moins, avec la valence syntaxique [33].

En parallèle, la linguistique slave, tournée vers les approches en dépendances, a également étudié les questions relatives aux alternances entre cadres de sous-catégorisation. C'est notamment le cas des écoles russe et tchèque. Citons à nouveau Meľčuk (1974) et Sgall *et al.* (1986). Le modèle développé par ce dernier, FGD, relie

33. Dans nos travaux sur la syntaxe lexicale (chapitre 5), et donc sur la sous-catégorisation syntaxique, les modèles restreints à la structure actancielle (sémantique) ou qui donnent la primauté au niveau sémantique ne sont donc pas directement pertinents ici. C'est le cas de FrameNet et des théories de l'alignement (*linking*) qui modélise la correspondance entre structure actancielle et sous-catégorisation syntaxique.

avant tout la valence au niveau dit tectogrammatical, c'est-à-dire approximativement au niveau sémantique : le lexique syntaxique est ainsi directement lié au lexique sémantique caractérisé par un inventaire d'actants typés au moyen de ce que l'on peut considérer quasiment comme des rôles thématiques (Agent, Patient, Effet, Direction…) (Panevová, 1994). C'est cette approche sur laquelle repose le Prague Dependency Treebank (Hajič *et al.*, 2006), ainsi que le lexique de valence VALLEX (Lopatková *et al.*, 2008). La TST, comme indiqué plus haut, a elle aussi donné naissance à un travail lexicographique important, dont une des composantes est la définition de ce que Milićević (2009) appelle *schéma de régime*, et qui correspond à un cadre de sous-catégorisation reliant, là aussi, le niveau sémantique et la réalisation syntaxique. Plus récemment, c'est à nouveau dans ce type de modèle que se place le développement du lexique de valence verbale du polonais Walenty (Przepiórkowski *et al.*, 2014), directement destiné à des applications de traitement automatique des langues.

Il se dégage de ce rapide survol la tendance suivante : le développement de ressources lexico-sytaxiques est une nécessité tant pour la linguistique descriptive que pour le traitement automatique des langues. Il s'agit d'un travail avant tout lexicographique, en ceci qu'il et nécessaire *in fine* de disposer d'informations précises pour chaque lexème, même si de nombreux travaux ont cherché à construire ou à aider à la construction de telles ressources (cf. section 5.2). Une ressource lexico-syntaxique ne peut se concevoir sans rapport avec la dimension sémantique, ne serait-ce que parce qu'une entrée lexicale, dans une telle ressource, doit correspondre à un sens spécifique de l'unité prédicative qu'elle décrit. En revanche, la façon de représenter l'information syntaxique et l'étroitesse de l'interaction entre structure argumentale syntaxique et structure actantielle sémantique varie d'une ressource et d'une approche à l'autre. Enfin, l'organisation du lexique en classes ou tables, ou au moyen de relations ou de règles, est une autre dimension de la variabilité des approches (ainsi les tables du Lexique-Grammaire ou les classes de Levin (1993)).

Il reste toutefois une double question importante : qu'est-ce qu'un argument syntaxique, et qu'est-ce qu'un argument sémantique, par opposition à la notion syntaxique de modifieur et à celle, sémantique, de circonstant ? C'est l'une des problématiques les plus débattues qui soient, et nous n'avons pas pour ambition de répondre à des questions aussi vastes. On peut par exemple se reporter à Boons *et al.* (1976b) et à Bonami (1999), qui montrent sur le français qu'il n'est pas vraiment possible de donner des critères clairs qui pourraient permettre de faire une distinction nette entre les deux. Nous verrons que cela se manifeste dans les ressources lexico-syntaxiques existantes par des choix divergents en la matière.

A.6 Développement de lexiques syntaxiques

Le développement de ressources lexico-syntaxiques a longtemps été le fait de linguistes et de lexicographes qui travaillaient principalement par introspection, éventuellement complétée par une exploration manuelle de corpus textuels. La raison en est double. Premièrement, l'analyse syntaxique de corpus, et singulièrement l'analyse automatique, ne permettait pas d'obtenir des résultats de qualité suffisante et sur des corpus de grande taille, ce qui interdisait toute tentative d'extraction automatique de ressources lexico-syntaxiques. Deuxièmement, le développement manuel de telles ressources était perçu comme le seul moyen de rassembler des informations riches sur des ensembles importants de lexèmes, y inclus les lexèmes les moins fréquents. En réalité, si la situation a changé concernant le premier de ces deux points, le deuxième, plus structurel, reste une question actuelle. Nous avons déjà mentionné l'entreprise de construction des tables du Lexique-Grammaire, initiée par Maurice Gross au Laboratoire d'Analyse Documentaire et Linguistique (LADL) (Gross, 1975 ; Boons *et al.*, 1976b,a ; Guillet et Leclère, 1992). Comme indiqué par Christian Leclère sur le site internet de l'équipe d'informatique linguistique de l'Institut Gaspard Monge (IGM, Université Paris-Est Marne-La-Vallée) [34], Maurice Gross s'attaque ainsi dès la fin des années 1960 à la question suivante soulevée par Chomsky (1965) : « Pour le moment, on peut à peine dépasser le simple arrangement classificatoire des données [lexicales]. Quant à savoir si ces limitations sont intrinsèques ou si une analyse plus profonde peut parvenir à débrouiller certaines de ces difficultés, cela reste en suspens. » Nous avons montré dans plusieurs travaux que rendre les tables du Lexique-Grammaire utilisables dans des outils de traitement automatique des langues a nécessité un travail conséquent (cf. chapitre 9). Toutefois, il s'agit là de la première entreprise de systématisation des propriétés syntaxiques, d'abord sur le français, puis sur plusieurs autre langues, au sein d'un réseau international de collaborations.

Le développement d'autres ressources lexico-syntaxiques pour le français a été initié à cette époque ou un peu plus tard, toujours avec une approche purement introspective. L'une d'entre elles est la ressource Les Verbes Français (LVF, Dubois et Dubois-Charlier, 1997), influencée par le Lexique-Grammaire mais s'en distinguant notamment par l'organisation des entrées en une structure hiérarchique de classes partageant des propriétés sémantiques homogènes [35]. De son côté, le lexique PROTON (van den Eynde et Mertens, 2003), développé à l'Université de Louvain à partir de 1986, et précurseur du lexique de valence verbal Dicovalence (van den Eynde et Mertens, 2006), est également le résultat d'un travail introspectif. Enfin, on peut citer le Dictionnaire Explicatif et Combinatoire, produit lexicographique de la Théorie Sens-Texte (TST, Meľčuk, 1974,

34. http ://infolingu.univ-mlv.fr/LADL/Historique.html [12.08.2014]

35. Ce qui ne veut pas dire que les classes sont homogènes sémantiquement. La majorité le sont, d'autres rassemblent par exemple des verbes dénominaux construits de façon comparable.

1988). Il est développé depuis plusieurs dizaines d'années (Meľčuk et Polguère, 1995 ; Meľčuk *et al.*, 1984, 1988, 1992, 1999), et a servi de précurseur au développement actuel du Réseau Lexical du Français (RLF, Lux-Pogodalla et Polguère, 2011), toujours dans le cadre de la TST. Le DEC, comme le RLF, est toutefois une ressource plus dictionnairique, qui ne comporte pas directement d'informations de sous-catégorisation, même si la structure actancielle (et donc sémantique) des éléments prédicatifs y est informée. Une source d'informations intéressante y est en revanche le riche inventaire de fonctions lexicales (des relations qui peuvent indiquer quels sont les verbes supports prototypiques pour un nom prédicatif, l'adjectif figé permettant de renforcer un nom — cf *peur bleue*) et d'autres informations de ce type qui ont une pertinence directe au niveau syntaxique. Le DEC, et pour l'instant encore le RLF, sont malheureusement des ressources dont la couverture reste très limitée, malgré la très haute qualité de leur contenu.

Pour l'anglais, les versions électroniques des dictionnaires pour apprenants restent longtemps la première source d'informations lexico-syntaxiques (Briscoe, 2001). C'est par exemple le cas des travaux menés à partir du Longman Dictionary of Contemporary English (LDOCE) (Boguraev et Briscoe, 1987 ; Boguraev *et al.*, 1987), qui ont donné naissance au lexique ANLT. Ce lexique, et d'autres ressources développées de cette façon, comme Comlex (Grishman *et al.*, 1994), ont une précision élevée de l'ordre de 95% mais un rappel relativement limité (76% pour ANLT, 84% pour Comlex), et ce malgré le travail manuel considérable qui a conduit à ces ressources.

Ni les tables du Lexique-Grammaire ni les lexiques ANLT et Comlex n'étaient librement acessibles à la communauté scientifique. Ces deux derniers lexiques, qui plus est, n'avaient pas pour entrées lexicales des lexèmes, mais bien des lemmes, ne permettant pas de départager les comportements syntaxiques distincts des différents lexèmes partageant un même lemme. Ce n'est pas le cas des tables du Lexique-Grammaire, qui distinguent des entrées différentes pour chaque emploi de chaque lemme. Mais le Lexique-Grammaire avait d'autres faiblesses, sur lesquelles nous reviendrons, dont les principales sont la mise à l'écart de classes de lexèmes verbaux problématiques mais très fréquents (les semi-auxilliaires, par exemple) et le caractère implicite des informations les plus importantes. En effet, le regroupement en classes d'entrées n'avait conduit qu'à renseigner explicitement les propriétés spécifiques distinguant chaque entrée des autres entrées de sa classe, mais pas les propriétés communes à toutes les entrées d'une même classe, informations que l'on ne pouvait trouver que dans les publications associées. Enfin, ces approches reposant directement ou indirectement (via des dictionnaires) sur l'introspection, elles ne peuvent modéliser les fréquence des emplois et des propriétés syntaxiques qui leur sont associées, et encore moins rendre compte des variations de ces fréquences d'un genre à l'autre, d'un domaine à l'autre, d'un style à l'autre, d'une époque à l'autre (Roland et Jurafsky, 1998).

Ce n'est qu'au début des années 1990 que les outils d'analyse automatique ont permis l'émergence d'une autre voie de recherche, dont l'ambition était de dépasser les cimites de l'approche lexicographique en cherchant à extraire des informations lexico-syntaxiques à partir de corpus, comme rappelé notamment par (Briscoe, 2001). On peut citer à cet égard les travaux de Brent (1991), puis de Manning (1993), qui exploite des analyses en chunks, ou de Ushioda *et al.* (1993), les premiers à extraire des informations fréquentielles, à partir d'une analyse morphosyntaxique. D'autres travaux ont suivi, et notamment ceux de Briscoe et Carroll (1997), qui s'appuient sur une analyse syntaxique complète. Des travaux plus récents font usage de techniques statistiques plus avancées, avec de bons résultats même sur des corpus analysés seulement morphosyntaxiquement (Lippincott *et al.*, 2012).

Les avantages de cette approche sont au moins triples (cf. Przepiórkowski, 2009 et références y incluses) : (i) un plus faible coût de développement de la ressource, (ii) un fondement empirique qui permet d'éviter les jugements variables voire discutables de lexicographes [36] et (iii) la possibilité d'extraire des informations de fréquences. Mais ces ressources extraites automatiquement ont plusieurs inconvénients majeurs : (i) leur précision est bien inférieure, d'une part en raison des erreurs des traitements appliqués comme l'étiquetage morphosyntaxique ou l'analyse syntaxique (Preiss *et al.*, 2007) et d'autre part en raison de la difficulté intrinsèque de la tâche — même pour les linguistes la distinction entre argument et modifieur n'est ni claire ni consensuelle —, (ii) la richesse des informations qu'elles contiennent est moins grande que dans les ressources obtenues par introspection, (iii) elles associent des informations syntaxiques à des lemmes et non à des lexèmes, et (iv) leur rappel sur corpus n'est pas plus satisfaisant que les ressources développées avec une approche lexicographique, il est vrai avec un coût bien plus élevé, en raison de la distribution fondamentalement zipfienne des cadres de sous-catégorisation en corpus (Korhonen *et al.*, 2000). De plus, elles reposent souvent, mais pas toujours, sur des inventaires de cadres de sous-catégorisation préexistants.

C'est également à partir des années 1990 qu'ont commencé à être développés des corpus arborés (*treebanks*), le premier d'entre eux étant le Penn TreeBank pour l'anglais (Marcus *et al.*, 1993), suivi par le Prague Dependency Treebank (Hajič *et al.*, 2000). Le travail théorique et le coût en termes d'annotation manuelle pour développer de telles ressources est très important, mais la disponibilité d'un corpus arboré ouvre la voie à l'extraction d'informations lexico-syntaxiques à partir des annotations qu'il contient, permettant ainsi la construction de ressources précises et munies d'informations fréquentielles. Toutefois, le problème du rappel se pose de façon encore plus cruciale, les corpus arborés étant

36. Ainsi, les tables du lexique-grammaire ont souvent tendance à considérer comme possibles pour une entrée donnée des constructions ou des propriétés syntaxiques conduisant à des énoncés proches de l'inacceptabilité, du seul fait que ces énoncés ne sont pas totalement inacceptables. Avec des implications souvent négatives dans des situations applicatives réelles (cf. section 9.1).

inévitablement de taille relativement modeste. De plus, les informations syntaxiques sont ici encore associées à des lemmes et non à des lexèmes, faute d'information sémantique associée aux mots dans le corpus arboré.

C'est ainsi que les trois approches principales mentionnées jusqu'ici, à savoir l'approche lexicographique (par introspection ou par exploitation de dictionnaires existants), l'exploitation de corpus annotés automatiquement et l'exploitation de corpus arborés, ont toutes été poursuivies jusqu'à ce jour. En voici quelques exemples :

- approche lexicographique :
  - le lexique Dicovalence de valence verbal du français, déjà cité,
  - PersPred, lexique de prédicats complexes du persan (Samvelian *et al.*, 2014),
  - plusieurs lexiques développés dans l'implémentation LKB du formalisme HPSG, notamment pour l'anglais, l'espagnol et l'allemand ;
  - plusieurs ressources importantes pour l'anglais qui se placent à l'interface entre cadres syntaxiques et valence sémantique, et notamment VerbNet (Kipper *et al.*, 2000), dont une version pour le français, Verbɜnet, est en cours de développement (Pradet *et al.*, 2014b), ainsi que PropBank (Palmer *et al.*, 2005) ;
- exploitation de corpus annotés automatiquement :
  - Valex, pour l'anglais (Korhonen *et al.*, 2006 ; Preiss *et al.*, 2007),
  - LexSchem pour le français (Messiant *et al.*, 2008),
  - Walenty pour le polonais (Przepiórkowski, 2009 ; Przepiórkowski *et al.*, 2014) ;
- exploitation de corpus arborés :
  - TreeLex, pour le français (Kupść, 2007, 2008) à partir du French TreeBank (Abeillé *et al.*, 2003),
  - lexique de valence pour le latin (McGillivray et Passarotti, 2009) extrait de l'*Index Thomisticus* Treebank (Passarotti, 2007),
  - lexique de valence pour le croate (Agić *et al.*, 2010) extrait du Croatian Dependency Treebank (Tadić, 2007),
  - VALLEX, pour le tchèque (Zabokrtský et Lopatková, 2007) à partir du Prague Dependency Treebank (Hajič *et al.*, 2006),
  - lexique de valence pour l'allemand (Hinrichs et Telljohann, 2009) associé au corpus arboré TüBa-D/Z (Telljohann *et al.*, 2009),

Ces deux dernières ressources sont du reste développées en parallèle avec le corpus arboré dont elles sont extraites, dans une dynamique d'enrichissement mutuel et simultané.

La raison pour laquelle ces trois grands types d'approches continuent à coexister vient de la complémentarité de leurs avantages et de leurs inconvénients, que nous

avons évoqués précédemment. Partir de corpus arborés permet d'extraire rapidement des informations de qualité, mais avec un rappel limité. Ceci suppose évidemment la disponibilité de tels corpus, dont le développement, pour leur part, est tout sauf rapide. Utiliser des techniques automatiques reste problématique quant à la qualité des informations extraites. Toutefois, nous avons pu montrer que des corpus analysés automatiquement peuvent non seulement produire de l'information lexicale, mais également contribuer à améliorer des informations lexicales existantes (cf. sections 5.2.2 et 9.1.3). Dans tous les cas, un travail manuel lexicographique reste toutefois indispensable pour garantir la richesse et la précision du lexique et pour en assurer une bonne couverture, y compris pour des cas rarement attestés en corpus.

Une autre limite de nombreuses ressources mentionnées ci-dessus est qu'elles se limitent souvent aux verbes simples. Or les verbes simples ne sont naturellement pas les seuls lexèmes susceptibles d'avoir des arguments syntaxiques. C'est par exemple le cas en français des adjectifs, des noms prédicatifs, de certains adverbes, et naturellement de nombreuses locutions et constructions à verbe support.

A.7 Développement automatique de wordnets

l'une des ressources sémantiques lexicales les plus connues et les plus utilisées dans les domaines du traitement automatique des langues et du web sémantique est le Princeton WordNet (PWN ; Fellbaum, 1998) et ses équivalents pour d'autres langues. Parmi ces derniers, on peut citer les wordnets développés dans le cadre des projets EuroWordNet (Vossen, 1999), BalkaNet (Tufiş, 2000) ou AsianWordnet (Sornlertlamvanich, 2010), ainsi que le Open Multilingual Wordnet (Bond et Paik, 2012), qui normalise et fusionne tous les wordnets dont la redistribution est autorisée par leurs auteurs, et inclut à ce jour des wordnets pour 27 langues. Initialement, le PWN était pourtant développé dans un contexte psycho-lexicographique (Miller, 1995), inspiré par des travaux sur les processus cognitifs d'accès au lexique.

Dans un wordnet, les lexèmes sont organisés en ensembles de synonymes, ou synsets, chaque synset représentant un sens. Un synset a un identifiant unique et contient donc un certain nombre de littéraux, qui sont approximativement des lemmes (simples ou composés), des termes voire des collocations, qui tous peuvent exprimer le sens représenté par le synset. Les synsets sont reliés entre eux par des relations sémantiques, la plus structurante étant la relation d'hypéronymie. Parmi les autres relations incluses dans le PWN on peut citer les relations de méronymie, d'holonymie ou d'antonymie. Par exemple, dans la version 3.1 du PWN, le synset nominal d'identifiant 02086723-n contient les littéraux {*dog*, *domestic dog*, *Canis familiaris*}. Le sens ainsi représenté est illustré par une définition (*a member of the genus Canis [...] that has been domesticated by man*

since prehistoric times ; occurs in many breeds) et un exemple d'emploi (*the dog barked all night*). Ce synset a deux hypéronymes, les synsets 02085998-n {*canine, canid*} 'canidé' et 01320032-n {*domestic animal, domesticated animal*}. Il a un certain nombre d'hyponymes, dont par exemple les synsets 02089774-n {*hunting dog*} et 02113929-n {*Newfoundland, Newfoundland dog*}.

Les premiers wordnets, et notamment le PWN, ont été développés manuellement, afin de maximiser la pertinence linguistique et de minimiser le taux d'erreur. Cependant, pour la grande majorité des langues, un tel effort est bien trop coûteux en temps et en moyens humains pour pouvoir être reproduit. C'est la raison pour laquelle diverses approches semi-automatiques et totalement automatiques ont été proposées pour le développement de wordnet à partir de divers types de ressources, et notamment en s'appuyant sur la disponibilité préalable du PWN.

Les techniques automatiques de développement de wordnets se répartissent selon celle des deux approches principales qu'elles mettent en œuvre : l'approche par fusion et l'approche par extension (Vossen, 1999). Dans le cas de l'approche par fusion, un wordnet pour une langue donnée est créé indépendamment des autres wordnets existants, en exploitant au mieux des ressources monolingues existantes ; dans un second temps, le wordnet ainsi créé peut être aligné avec les wordnets disponibles pour d'autres langues (Rudnicka *et al.*, 2012). Dans le cas de l'approche par extension, que nous avons utilisée, l'inventaire de sens du PWN est conservé (mêmes identifiants de synsets, mêmes relations entre synsets) et on cherche à peupler les synsets avec des littéraux de la langue cible, par exemple par désambiguïsation et traduction des littéraux anglais présents dans le PWN.

L'approche par extension repose donc sur l'approximation selon laquelle les concepts (les sens) et les relations entre eux sont indépendants de la langue, au moins pour une bonne part. C'est du reste la principale limite de cette approche : les wordnets produits sont biaisés par rapport au PWN ce qui peut même, dans certains cas, rendre certains synsets ou certaines relations arbitraires (Orav et Vider, 2004 ; Wong, 2004). Par exemple, le PWN contient un synset {*performer, performing artist*}, défini comme étant *un artiste réalisant un spectacle théâtral ou musical devant face à une audience.* Mais il n'y a de mot ni en français ni en slovène qui dénote de façon globale les acteurs, chanteurs et autres artistes se produisant en spectacle. Dans un tel cas, il est toujours possible de laisser vide le synset en question dans la ressource produite. À l'inverse, certains sens raisonnablement répandus de la langue cible peuvent ne pas correspondre à un synset du PWN, par exemple parce qu'ils n'ont pas vraiment de réalité culturelle dans les pays de langue anglaise ou ne sont pas considérés comme suffisamment importants. C'est ainsi le cas des sens ou concepts dénotés en français par *raclette*, *Jacques Chirac* ou *École Polytechnique*. Parfois, c'est le découpage même en synsets qui ne correspond pas bien. Ainsi, les synsets {*lawyer, attorney*} (*a professional person authorized to practice law ; conducts lawsuits or gives legal*

advice) et {*advocate, counsel, counselor, counsellor, counselor-at-law, pleader*} (*a lawyer who pleads cases in court*) sont distingués selon des critères propres au système judiciaire anglo-saxon voire américain, qui ne se superposent pas avec les distinctions françaises entre juriste, avocat ou avoué.

Malgré tout, ces problèmes sont plus que compensés par les avantages importants de l'approche par extension, qui est ainsi très utilisée pour le développement de wordnets, par exemple dans les projets BalkaNet (Tufiş, 2000), MultiWordnet (Pianta *et al.*, 2004) et BabelNet (Navigli et Ponzetto, 2010). Le premier avantage est naturellement un coût de développement bien plus bas que pour l'approche par fusion. Le second avantage est que les wordnets produits sont alignés sur le PWN et donc également alignés entre eux, ce qui permet d'envisager leur utilisation dans des applications multilingues, telles que la traduction automatique ou l'extraction d'informations. Nous ne prétendons pas que l'approche par extension soit meilleure que l'approche par fusion, mais nous pensons qu'il s'agit du meilleur choix dans un contexte où, comme dans le nôtre, on souhaite développer à moindre coût une ressource à large couverture et de précision suffisante pour la majorité des applications possibles.

Les mises en œuvre de l'approche par extension varient selon le type de ressources qui sont disponibles pour la construction d'un wordnet dans une langue donnée. Les premiers travaux en ce sens utilisaient directement des dictionnaires électroniques bilingues, en essayant d'aligner les entrées des dictionnaires avec wordnet (Knight et Luk, 1994 ; Yokoi, 1995). Le problème qui apparaît immédiatement est la difficulté de la désambiguïsation des (sous-)entrées des dictionnaires bilingues. De plus, les dictionnaires bilingues ont souvent une couverture limitée et ne sont pas nécessairement disponibles pour toutes les paires de langues.

Une façon de surmonter ces difficultés est d'utiliser des lexiques bilingues ou multilingues extraits de corpus parallèles (Resnik et Yarowsky, 1997 ; Fung, 1995). L'hypothèse principale qui sous-tend ces travaux est que les différents sens d'un même mot qui est ambigu dans une langue sont souvent traduits par des mots différents dans une autre langue. De plus, si deux mots distincts, voire plus, sont traduits par le même mot dans une autre langue, ils sont souvent sémantiquement reliés, voire synonymes. Ceci permet de désambiguïser les mots polysémiques ou à l'inverse de créer des liens synonymiques. Les corpus parallèles ont été utilisés pour induire des synsets pour une nouvelle langue par divers auteurs (Dyvik, 2002 ; Ide *et al.*, 2002 ; Diab, 2004).

La troisième famille de travaux, plus récente, cherche à exploiter au mieux les ressources libres et collaboratives telles que Wikipedia. Wikipedia est une ressource encyclopédique libre disponible dans de nombreuses langues. Chaque article peut notamment être muni de catégories et être relié à des articles qui lui correspondent dans les Wikipedia d'autres langues. De nouveaux wordnets ont été induits en associant des

pages de Wikipedia avec des synsets de wordnet (Suchanek *et al.*, 2008), en utilisant des informations structurelles pour associer des catégories Wikipedia aux synsets (Ponzetto et Navigli, 2009) ou en extrayant des mots-clés à partir des articles de Wikipedia (Reiter *et al.*, 2008). Un modèle vectoriel permettant d'associer des pages Wikipedia à wordnet a été proposé par divers auteurs (Ruiz-Casado *et al.*, 2005 ; Declerck *et al.*, 2006). Les approches les plus abouties utilisent Wikipedia et d'autres ressources wiki, notamment Wiktionary (version française : le Wiktionnaire) pour produire des wordnets dans de multiples langues (de Melo et Weikum, 2009 ; Navigli et Ponzetto, 2010, 2012).

A.8 Aperçu historique des travaux académiques en correction orthographique

Les travaux sur la correction lexicale ont débuté dans les années 1960 (Blair, 1960 ; Damerau, 1964). Au fil du temps et des avancées technologiques, ce domaine de recherche a beaucoup évolué (Kukich, 1992 ; Mitton, 1996, 2010). Les premiers systèmes proposés se limitaient au mot à corriger sans prendre en compte leur contexte d'apparition. Ils fonctionnaient principalement à base de règles typographiques et de distance d'édition (Damerau, 1964 ; Kernighan *et al.*, 1990) ou avec un système de vérification dans le lexique plus tolérant (Oflazer, 1996). Puis le contexte a commencé à être pris en compte. Pour ce faire, certains travaux se sont appuyés sur des modèles de langue *n*-grammes de mots (Brill et Moore, 2000 ; Carlson et Fette, 2007 ; Park et Levy, 2011), phonétiques (Toutanova et Moore, 2002), ou encore sur des modèles de langue *n*-grammes qui combinaient ces deux caractéristiques (Boyd, 2009). Ces *n*-grammes sont souvent associés par la suite à d'autres paramètres tels que la catégorie grammaticale, la transcription phonétique ou encore la longueur du mot à corriger. D'autres auteurs ont proposé d'utiliser des mesures distributionnelles (Suignard et Kerroua, 2013) ou sur des modèles probabilistes (Yvon, 2011).

Le choix des techniques utilisées est souvent guidé par l'objectif visé par le correcteur. Un correcteur qui cherche à corriger des fautes grammaticales en plus des fautes lexicales (cf. Carlson et Fette, 2007 ; Yvon, 2011) doit être plus robuste, la détection d'erreurs étant plus complexe à réaliser et la génération de candidats de correction plus risquée. Par ailleurs, si le correcteur en question doit traiter des textes plus dégradés (SMS, forum, blogs, etc.), les approches doivent être adaptées, par exemple en faisant usage de lexiques de correction spécifiques (cf. ci-dessous section 7.3.3, mais aussi Guimier de Neef *et al.*, 2007), en phonétisant le texte à traiter (Kobus *et al.*, 2008) ou en recourant à un apprentissage par alignement (Beaufort *et al.*, 2010).

Les systèmes qui se concentrent uniquement sur les fautes lexicales, altérations dont le résultat est un token inconnu d'un lexique de référence, reposent quant à eux souvent sur

la notion de distance d'édition (Damerau, 1964 ; Levenshtein, 1965, 1966). Cette notion met en œuvre quatre types de règles (l'insertion et la suppression d'une lettre, la substitution d'une lettre par une autre, l'inversion de deux lettres consécutives). L'idée est de s'appuyer sur ces opérations, que l'on peut donc considérer comme des règles de correction, pour passer d'un mot mal orthographié à sa forme attendue. Dans nos travaux sur l'outil TEXT2DAG (Sagot et Boullier, 2008 ; cf. section 7.3.1), nous avons étendu ou adapté cette méthode pour la modélisation explicite des fautes de proximité clavier et des fautes phonétiques (par exemple, le remplacement de `o` par `eau` ; cf. aussi Véronis, 1988). Le contexte dans lequel s'applique la règle peut être pris en compte (Kernighan *et al.*, 1990, cf. aussi section 7.3.2), jusqu'à l'utilisation de règles spécifiques lorsque l'on cherche à corriger des segments spécifiques comme des entités nommées de tel ou tel type (Gábor et Sagot, 2014 ; Sagot et Gábor, 2014, cf. section 7.3.4). Par ailleurs, la pondération de ces règles ne se fait plus systématiquement en fonction du nombre d'opérations effectuées sur le mot. Le choix du poids d'une correction se fait, manuellement ou non, en fonction de l'opération effectuée et des lettres concernées par cette opération (Véronis, 1988 ; Mitton, 1996 ; Sagot et Boullier, 2008) voire par apprentissage supervisé (Baranes et Sagot, 2014b ; Baranes, 2015, cf. section 7.3.2).

A.9 Analyse morphosyntaxique

L'étiquetage morphosyntaxique (*tagging*) est une tâche désormais classique en traitement automatique des langues, pour laquelle de nombreux systèmes ont été développés ou adaptés à un large éventail de langues. Elle consiste à associer à chaque « mot » une *étiquette morphosyntaxique* dont la granularité peut aller d'une simple catégorie morphosyntaxique, ou partie du discours, à une catégorie plus fine et enrichie par des traits morphologiques (genre, nombre, cas, temps, mode, etc.). Les premiers systèmes d'étiquetage morphosyntaxique étaient des systèmes pour l'anglais à base de règles. Le premier d'entre eux, développé en 1958–59, est probablement l'étiqueteur intégré au *Transformations and Discourse Analysis Project* de Harris (1962). Il reposait uniquement sur un dictionnaire d'étiquettes et un jeu de 14 règles de désambiguïsation développées à la main et implémentées sous une forme qui préfigure les cascades de transducteurs (Joshi et Hopely, 1996). Peu après vient ensuite l'étiqueteur de Klein et Simmons (1963), un système simple reposant là aussi sur des règles de désambiguïsation, mais également sur des listes de suffixes et un lexique de 400 mots outils et de 1 500 mots supplémentaires qui font exception aux règles. La précision est de l'ordre de 90%, avec un jeu de 30 étiquettes. Un peu plus tard a été développé TAGGIT (Greene et Rubin, 1971), étiqueteur destiné à la pré-annotation en étiquettes morphosyntaxiques du Brown Corpus (Francis et Kucera, 1964). Le jeu d'étiquettes est plus précis (86 étiquettes), et il repose

sur environ 3 300 règles, près de 500 suffixes et un lexique d'environ 3 000 exceptions. La précision sur le Brown Corpus est estimée à 77%, ce score inférieur s'expliquant par la granularité plus fine du jeu d'étiquettes. C'est avec les outils ayant servi au développement du Lancaster–Oslo/Bergen Corpus (LOB, Johansson *et al.*, 1978) que sont mises au premier plan des approches statistiques pour l'étiquetage morphosyntaxique (Marshall, 1983). Cet étiqueteur, nommé CLAWS par la suite (Booth, 1985 ; Marshall, 1987), repose en partie sur une approche similaire à celle de TAGGIT, quoiqu'avec plus de règles, un lexique plus grand, et un jeu d'étiquettes plus large (130 étiquettes). Mais il repose également sur l'utilisation de probabilités bigrammes extraites du Brown Corpus. Une fois complété par l'algorithme IDIOMTAG (Leech *et al.*, 1983), que l'on pourrait aujourd'hui qualifier de système de détection et de ré-étiquetage de composés, la précision du système est estimée comme étant située entre 96% et 97%, au prix d'un travail manuel considérable et impossible à transférer tel quel à une autre langue. Il est néanmoins intéressant de constater que l'ordre de grandeur des performances des meilleurs systèmes d'étiquetage morphosyntaxique de l'anglais était déjà si élevé il y a une trentaine d'années.

L'utilisation de probabilités bigrammes est un exemple simple d'un paradigme d'étiquetage statistique développé à partir du milieu des années 1980 : l'utilisation de modèles de Markov cachés bigrammes puis trigrammes, souvent implémentés au moyen d'algorithmes de programmation dynamique (Church, 1988). Au milieu des années 1990, ces systèmes étaient considérés comme état-de-l'art (Merialdo, 1994 ; Brants, 1996), comme précisé par Kim *et al.* (1999) dans leur article sur l'introduction limitée d'informations lexicales dans un tel modèle. D'autres approches statistiques ont été mises en œuvre également à cette époque, et notamment les arbres de décision (Schmid, 1994 ; Magerman, 1995) et les modèles à maximisation d'entropie (Maximum Entropy Markov Models, MEMM ; Ratnaparkhi, 1996 ; cf. aussi le chapitre 8). De plus, l'arrivée du PTB (Marcus *et al.*, 1993 ; cf. aussi la section A.6) a permis aux systèmes de se comparer entre eux de façon précise, l'état de l'art étant alors de 96,5% environ. Pour un panorama plus complet des travaux sur l'étiquetage morphosyntaxique avant les années 2000, on pourra se reporter à Manning et Schütze (1999).

Depuis, les étiqueteurs morphosyntaxiques pour l'anglais les plus performants reposent presque tous, à la suite de l'étiqueteur de Ratnaparkhi (1996), sur des modèles linéaires discriminants tels que les modèles à maximisation d'entropie (Toutanova et Manning, 2000) ou les perceptrons moyennés, avec différents types de raffinements. D'autres approches ont toutefois continué à être explorées, et notamment les modèles de Markov cachés (étiqueteur TnT, Brants, 2000) ou les arbres de décision. L'étiqueteur le plus performant actuellement sur le PTB, le Stanford Tagger dans sa version 2.0 (Manning, 2011), atteint 97,29% de précision, mais seulement 89,70% de précision sur les seuls

mots inconnus [37]. Il est interessant de noter que (Manning, 2011) considère qu'on ne peut attendre à ce niveau de précision qu'une amélioration « limitée » via un « meilleur apprentissage automatique ou de meilleurs traits dans un système séquentiel discriminant » [38], et que « les perspectives de gains supplémentaires par apprentissage semi-supervisé semblent également très limitées » [39]. Pour lui, les problèmes linguistiques sous-jacents à la tâche d'étiquetage morphosyntaxique telle qu'elle est définie et mise en œuvre dans l'annotation des corpus font que l'on a vraisemblablement atteint un plafond difficile à dépasser. C'était du reste déjà l'avis de Ratnaparkhi (1996), qui affirmait il y a près de vingt ans que tout algorithme reposant sur un corpus [a peu de chances] d'atteindre sur le [PTB] des performances bien supérieures à 96,5% en raison de problèmes de cohérence » [40].

Pour le français, ces différentes approches ont donné lieu au développement ou à l'adaptation de plusieurs étiqueteurs morphosyntaxiques. Outre notre étiqueteur MElt (Denis et Sagot, 2009, 2010, 2012), dont il est question au chapitre 8, on peut notamment citer les outils suivants, dont certains sont apparus plus récemment :

— TreeTagger [41], qui repose sur l'apprentissage d'arbres de décision (Schmid, 1994) et qui, avec le modèle fourni pour le français, constitue probablement l'étiqueteur le plus utilisé dans la communauté francophone, certainement en raison de sa gratuité [42], de son ancienneté et de sa facilité d'utilisation, malgré des performances inférieures à plusieurs étiqueteurs plus récents ;

— LIA_tagg, étiqueteur librement disponible qui implémente un modèle de Markov caché (Nasr *et al.*, 2004) ;

— TnT, étiqueteur disponible sous une licence plus restrictive que TreeTagger et qui repose également sur un modèle de Markov caché (Brants, 2000) ;

— l'adaptation au français de l'étiqueteur de Stanford (Toutanova et Manning, 2000 ; Manning, 2011), qui repose sur un modèle de Markov à maximisation d'entropie, dont l'originalité est de procéder de façon bidirectionnelle ;

37. Le wiki de l'ACL propose un tableau récapitulatif des caractéristiques et des performances des meilleurs étiqueteurs morphosyntaxiques pour l'anglais (et le français) à l'adresse suivante : `http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)`. On y trouvera des scores plus élevés que les 97,29% du Stanford Tagger, mais ils sont le fait de systèmes non disponibles et/ou dont l'apprentissage repose sur des données textuelles supplémentaires via des approches faisant usage d'apprentissage semi-supervisé, d'auto-apprentissage ou de calculs de similarité distributionnelle.

38. « there is limited further mileage to be had either from better machine learning or better features in a discriminative sequence classifier. »

39. « The prospects for further gains from semisupervised learning also seem quite limited. »

40. La phrase complète est la suivante : « The convergence of the accuracy rate implies that either all these techniques are missing the right predictors in their representation to get the "residue", or more likely, that any corpus based algorithm on the Penn Treebank Wall St. Journal corpus will not perform much higher than 96.5% due to consistency problems. »

41. Disponible sur `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`.

42. TreeTagger est un outil gratuit mais pas libre, puisque son code source n'est pas disponible.

— LGtagger, étiqueteur récent et lui aussi librement disponible, qui repose sur les champs aléatoires conditionnels (Conditional Random Fields, CRF, (Lafferty *et al.*, 2001) ; sa spécificité est de traiter comme une tâche jointe à la fois la reconnaissance des mots composés et l'étiquetage morphosyntaxique, en s'appuyant notamment sur des ressources lexicales externes, dont le Le*fff* (Constant *et al.*, 2011 ; Constant et Sigogne, 2011).

Ces étiqueteurs obtiennent sur le FTB (Abeillé *et al.*, 2003) des performances allant de 96,1% à 97,8% (cf. plus bas). On notera que certains articles (cf. par exemple Falk *et al.*, 2014) comparent les performances de tels systèmes à celles obtenues par des analyseurs syntaxiques complets — une fois supprimées les structures syntaxiques produites, pour ne conserver que les étiquettes morphosyntaxiques associées aux mots par l'analyseur. Ce sont là des comparaisons inéquitables. En effet, le développement d'un analyseur syntaxique nécessite des modèles linguistiques ou des corpus annotés ainsi que des algorithmes d'analyse bien plus complexes que ce qui relève directement de l'étiquetage morphosyntaxique.

Pour certaines langues telles que l'anglais et quelques autres langues très étudiées dont le français, les systèmes d'étiquetage morphosyntaxique statistique ont ainsi atteint des niveaux de performances difficiles à dépasser, et qui ne sont d'ailleurs plus très éloignés des performances d'un annotateur humain, du moins quantitativement — les erreurs effectuées par les humains et par les outils automatiques ne sont pas toujours les mêmes. Curieusement, la plupart des systèmes d'étiquetage morphosyntaxique ne reposent pas sur des sources d'informations lexicales externes de grande taille, mais plutôt sur un lexique extrait du corpus utilisé pour l'apprentissage (cf. cependant Hajič, 2000). Ceci nous a amené à travailler sur les questions suivantes : une intégration raisonnée de connaissances lexicales externes dans un système d'étiquetage morphosyntaxique permet-elle d'en améliorer les performances ? Quel type d'intégration est-il ici optimal, par exemple sous forme de contraintes (restriction de l'inventaire d'étiquettes possibles pour un mot donné) ou de traits ? Quelles informations lexicales externes convient-il d'utiliser ? À l'évidence, un des avantages qu'il pourrait y avoir à utiliser des connaissances lexicales externe est que cela devrait permettre un meilleur traitement des mots inconnus du corpus d'apprentissage, pour peu qu'ils soient présents dans la ressource lexicale utilisée. Il y a là peut-être une piste pour améliorer le comportement d'un étiqueteur morphosyntaxique sur des types de données qui ne sont pas du même genre, du même niveau de langue, de la même variété géographique ou de la même époque que les données d'apprentissage utilisées. De plus, dans le cas d'un modèle d'étiquetage séquentiel étiquetant de gauche à droite, comme la plupart des modèles d'étiquetage, une ressource lexicale externe peut permettre d'avoir accès à des informations, même non désambiguïsées, concernant le contexte droit du mot à étiqueter.

L'utilisation de techniques d'apprentissage automatique présuppose la disponibilité de ressources annotées sur lesquelles s'entraîner. Or le développement de telles ressources est coûteux. Le développement de lexiques est également une tâche coûteuse, mais nous avons vu précédemment qu'il était possible de l'accélerer au moyen de techniques dédiées. La question se pose alors de savoir si les améliorations obtenues grâce à l'utilisation d'un lexique externe permettent de réduire le coût de développement des ressources nécessaires à la construction d'un étiqueteur morphosyntaxique. Autrement dit, à niveau de qualité identique pour les étiqueteurs morphosyntaxiques construits *in fine*, est-il possible d'une part et moins coûteux en temps d'autre part de réduire la taille du corpus d'apprentissage nécessaire grâce au développement en parallèle d'un lexique externe ?

A.10 Analyse syntaxique

A.10.1 Analyse syntaxique symbolique et analyse syntaxique probabiliste : un bref historique

En dehors de la traduction automatique, l'analyse syntaxique est probablement le plus ancien des domaines de recherche en traitement automatique des langues. Il s'est longtemps restreint à des approches symboliques, reposant sur des grammaires écrites à la main. On le fait souvent remonter à Yngve (1955), qui décrit ce qui semble être le premier algorithme d'analyse syntaxique ascendant (*bottom-up*), destiné à être intégré dans une architecture de traduction automatique — sans que le résultat ne soit convaincant. Des algorithmes descendants (*top-down*) ont été alors rapidement proposés, dont le *Multiple-Path Syntactic Analyzer* de Kuno et Oettinger (1963), qui repose sur une grammaire non contextuelle (*Context-Free Grammar*, CFG) d'environ 3 400 règles de réécriture et sur un inventaire de catégories relativement fin [43]. C'est de fait autour de l'analyse au moyen de grammaires non contextuelles que s'est d'abord concentrée la majorité des efforts. L'analyse syntaxique non contextuelle, avant tout développée dans le contexte de la compilation de langages de programmation de plus en plus évolués (ALGOL68, Pascal) au moyen de grammaires déterministes (Aho et Ullman, 1972 ; Hopcroft et Ullman, 1979 ; Aho *et al.*, 1986 et la version française de sa seconde édition, Aho *et al.*, 2007), a pris un essor important grâce au traitement automatique des langues. En effet, ces dernières se distinguent des langages de programmation par de nombreux aspects, dont leur très forte ambiguïté qui se manifeste par le besoin d'algorithmes d'analyse non déterministes. Le premier algorithme de ce type, l'algorithme CKY, est proposé au début des années 1960. C'est le résultat du travail non publié de Cocke ultérieurement raffiné par Kasami (1965) et Younger (1967). Cet algorithme suppose que la grammaire est en forme normale de

43. Par exemple, le mot anglais *are* est noté comme ambigu entre trois catégories : verbe intransitif, copule et auxiliaire.

Chomsky. Earley (1968, 1970) propose alors un algorithme, lui aussi en temps cubique, qui peut s'appliquer à toute grammaire non contextuelle. Naturellement, le nombre d'analyses possibles pour une phrase donnée avec une grammaire non contextuelle générale pouvant être exponentiel, cette complexité cubique est obtenue par l'application de techniques de programmation dynamique, et notamment l'utilisation de forêts partagées qui permettent de représenter de façon factorisée l'ensemble des analyses. Ces algorithmes, dont Lang (1974) [44] propose une analyse unifiée et démontre nombre de propriétés, sont des cas particuliers de ce que Kay (1980) nommera l'analyse par chartes (*chart parsing*).

Comme nous l'avons vu à la section A.5, la fin des années 1970 et le début des années 1980 voient l'arrivée de modèles syntaxiques lexicalisés, en réaction aux grammaires transformationnelles dont les propriétés avaient découragé la communauté de l'analyse syntaxique des langues (Peters et Ritchie, 1973). Certains de ces nouveaux formalismes, comme les Grammaires Lexicales Fonctionnelles (*Lexical Functional Grammar*, LFG ; Bresnan, 1982), sont explicitement construites autour d'un formalisme squelette, les grammaires non contextuelles, qui modélisent la constituance, et de décorations structurées que l'on manipule par des opérations d'unification. Ce sont donc des formalismes à deux niveaux (Sagot, 2006). Les décorations permettent de modéliser plus finement les mécanismes syntaxiques à l'œuvre, et apportent une réponse à la prise de conscience que les grammaires non contextuelles ne sont pas suffisantes pour modéliser les langues (Shieber, 1987). Elles ont l'inconvénient, dans le cas général, d'induire pour les analyseurs une complexité en temps potentiellement exponentielle, mais ont l'avantage de permettre la prise en compte d'informations riches, y compris de manière non locale (par exemple pour vérifier la conformité d'une analyse à des informations de sous-catégorisation, y compris en présence d'extractions ou d'autres phénomènes non locaux). Les informations lexicales, y compris syntaxiques, y jouent ainsi un rôle central. L'autre piste permettant d'augmenter l'expressivité des formalismes en jeu consiste à augmenter la puissance d'expression du formalisme squelette. C'est le cas par exemple avec la définition des grammaires d'adjonction d'arbres (*Tree Adjoining Grammars*, TAG, Joshi *et al.*, 1975 ; Joshi, 1987), voire de formalismes plus expressifs encore, tout en restant dans le champ des formalismes analysables en temps polynomial [45].

44. Version étendue d'un précédent manuscrit de 1970.

45. Au-delà du niveau d'expressivité des TAG, on peut citer deux niveaux remarquables : (i) celui des systèmes de réécriture linéaire non contextuels (*Linear Context-Free Rewriting Systems*, LCFRS ; Vijay-Shanker *et al.*, 1987), dont la puissance d'expression est identique à celle des TAG multi-composants (MC-TAG ; Joshi *et al.*, 1975 ; Joshi, 1987) et à certaines de leurs variantes, comme les Grammaires d'Insertion d'Arbres (TIG) multi-composants (MC-TIG ; Boullier et Sagot, 2009a) et (ii) celui, qui en est un sur-ensemble, correspondant à l'ensemble des langages analysables en temps polynomial, PTIME. Ce dernier niveau est exactement celui des Grammaires à Concaténation d'Intervalles (*Range Concatenation Grammars*, RCG ; Boullier, 1999, 2000) pour lesquelles existent des analyseurs efficaces, et notamment celui développé par Pierre Boullier et intégré à l'environnement Syntax (Barthélemy *et al.*, 2001). J'ai montré pendant ma thèse que la non-linéarité des RCG pouvait être intéressante pour la modélisation du langage (Sagot et Boullier, 2004) et comment cette idée pouvait être mise en œuvre (Sagot, 2005b). Pierre Boullier et moi-même nous sommes penchés depuis

Ces deux pistes ne sont pas incompatibles, et les TAG aussi ont servi de squelette à des formalismes à deux niveaux dont les décorations sont là aussi calculées par unification (Shieber, 1986). Depuis, des systèmes d'analyse syntaxique symbolique à grande échelle ont été développés. Ils reposent souvent sur l'un des formalismes lexicalisés majeurs (LFG, HPSG, TAG, CCG), mais également sur des implémentations efficaces d'algorithmes utilisant la programmation dynamique [46]. Ils construisent généralement des forêts partagées d'analyse à partir du formalisme squelette, structures qui sont filtrées (au cours de leur construction ou dans une phase ultérieure) grâce au calcul des décorations. Il faut naturellement développer la grammaire et le lexique à la main, souvent à l'aide d'environnements dédiés qui permettent l'édition de la grammaire et sa compilation en un analyseur syntaxique. Il en est ainsi de LKB pour HPSG (Copestake et Flickinger, 2000), du *Grammar's Writer Workbench* (Kaplan et John T. Maxwell, 1993–1996) et de XLE (Butt et King, 2003) pour LFG ou de XMG (Crabbé *et al.*, 2013) et de l'écosystème DyALog/MgComp (Thomasset et Villemonte de La Clergerie, 2005) pour les TAG (à travers un niveau d'abstraction supplémentaire, celui des métagrammaires au sens de Candito, 1999). C'est dans ce dernier environnement qu'a été notamment développé FRMG, grammaire et analyseur syntaxique pour le français.

En parallèle à ces développements, les approches statistiques pour l'analyse syntaxique ont émergé pendant les années 1990 grâce à la disponibilité pour l'anglais du Penn Treebank (PTB, cf. section A.6). Ce corpus arboré étant annoté en constituants, ce type d'analyse est resté l'objet d'une large majorité des efforts, par-delà le passage d'un paradigme symbolique à un paradigme statistique. L'avantage principal de ce dernier est de permettre la représentation dans un même formalisme des informations nécessaires à l'analyse proprement dite et à la désambiguïsation, informations qui sont extraites à partir d'un corpus arboré. Bien que, le PTB ait été suivi par d'autres corpus arborés développés pour d'autres langues, comme rappelé à la section A.5, les méthodes d'analyse syntaxique statistique ont été développées avant tout à partir du PTB, et donc orientées vers l'analyse en constituants de l'anglais, langue typologiquement peu représentative [47]. Les premiers modèles utilisés étaient des modèles génératifs qui reposaient sur des grammaires non

sur la problématique de l'analyse de DAG d'entrée avec des RCG, rendue délicate par la non-linéarité des RCG (Boullier et Sagot, 2009b).

46. Nous n'évoquerons pas ici notre analyseur SxLfg, analyseur reposant sur une grammaire LFG produite par une grammaire écrite dans un formalisme plus abstrait (Boullier et Sagot, 2005 ; Boullier *et al.*, 2005b ; Sagot, 2006 ; Sagot et Boullier, 2006) et compilée en un analyseur tabulaire par une version étendue à LFG du générateur d'analyseurs Syntax (Boullier et Deschamp, 1988–2007).

47. Ce que nous mettons ici en avant est le caractère fortement configurationnel de l'anglais, qui fait écho à la simplicité de sa morphologie flexionnelle. À l'échelle des langues du monde, et même à celle des seules langues européennes, l'anglais, de par ces caractéristiques, est très loin d'être une langue « moyenne » ou « médiane », quoi que cela puisse vouloir dire. Nous évoquons brièvement à la section A.10.2 sur les efforts récents pour fédérer les chercheurs travaillant sur l'analyse syntaxique de langues typologiquement moins atypiques que l'anglais, notamment par leur morphologie flexionnelle plus riche et plus informative ainsi que par leur ordre des mots souvent moins fixe.

contextuelles probabilisées (*Probabilistic Context-Free Grammars*, PCFG) extraites à partir du PTB (Charniak, 1997 ; Collins, 1997 ; Johnson, 1998 ; Klein et Manning, 2003 ; Collins, 2003 ; Petrov *et al.*, 2006 ; McClosky *et al.*, 2006). On peut considérer les PCFG comme un autre type de formalisme à deux niveaux, avec un squelette non contextuel, à l'expressivité par conséquent limitée, et des décorations qui sont des nombres réels — les probabilités — calculés par multiplication (et non des structures calculées par unification comme par exemple en LFG). Les premières approches reposant sur les PCFG souffraient toutefois de deux limitations importantes : (i) les symboles non terminaux y sont souvent insuffisamment spécifiques, et l'hypothèse de non-contextualité, ou hypothèse d'indépendance, est alors trop forte, et (ii) l'information lexicale y est sous-utilisée. Depuis une dizaine d'années, différents travaux ont été publiés qui cherchent à dépasser ces limitations, atteignant ainsi des niveaux élevés de performances. On peut classer ces modèles d'analyse en deux grandes familles :

— les modèles lexicalisés, qui utilisent directement les mots (cf. plus bas) comme symboles terminaux et/ou qui font percoler de bas en haut dans l'arbre de constituants les têtes de chaque syntagme (Collins, 2003) ;

— les modèles non lexicalisés, qui utilisent les étiquettes morphosyntaxiques comme terminaux. On peut notamment citer l'algorithme de Petrov *et al.* (2006) puis Petrov et Klein (2007), utilisé dans l'analyseur de Berkeley mais aussi dans LORG (Attia *et al.*, 2010). Il permet de rétablir la pertinence de l'approximation de non-contextualité grâce au raffinement itératif automatique des symboles de la grammaire, idée mise en œuvre manuellement par Klein et Manning (2003) : cela permet de modéliser plus finement les comportements syntaxiques de certains types de mots ou de constituants partageant une même étiquette (respectivement morphosyntaxique ou syntaxique).

On constate également l'émergence d'analyseurs probabilistes reposant sur des squelettes plus expressifs que les grammaires non contextuelles, faisant ainsi le lien avec les résultats obtenus par le passé, avant la généralisation des approches probabilistes. C'est ainsi qu'ont été définies et mises en œuvre les TAG probabilisées (Resnik, 1992 ; Chiang, 2000), puis, plus récemment, des formalismes probabilisés reposant sur un squelette encore plus expressif (Plaehn, 2005 ; Kallmeyer et Maier, 2010).

Les tâches partagées (*shared tasks*) sur l'analyse syntaxique en dépendances (Buchholz et Marsi, 2006 ; Nivre *et al.*, 2007a) ont néanmoins montré l'intérêt de l'analyse syntaxique statistique en dépendances, à côté de l'analyse en constituants. On peut notamment citer, sur le plan de la modélisation syntaxique, la possibilité de représenter et donc de construire automatiquement des structures non projectives, ainsi que la disponibilité de corpus arborés annotés en dépendances (par exemple le Prague Dependency Treebank ; Hajič *et al.*, 2006). De plus, c'est là qu'ont émergé les approches discriminantes pour

la construction des structures syntaxiques, par opposition aux approches purement génératives ou à celles faisant usage d'une approche discriminante pour réordonner les analyses concurrentes proposées par un analyseur génératif (Collins, 2000 ; Charniak et Johnson, 2005). On peut notamment citer les deux principales approches pour l'analyse syntaxique statistique en dépendances que sont :

— l'analyse par transitions, représentée notamment par l'analyseur MaltParser (Nivre *et al.*, 2006, 2007b),

— l'analyse par graphes, dont l'approche la plus classique est représentée par l'analyseur MSTParser (McDonald *et al.*, 2005) et l'analyseur MATE de Bohnet (2010) qui en est une amélioration (cf. section 9.3) ; se rangent également dans cette catégorie, quoi que de façon différente, l'analyseur TurboParser de Martins *et al.* (2010, 2013) ou encore l'analyseur de Huang et Sagae (2010).

Outre d'excellentes performances, ces approches ont permis le développement d'analyseurs très efficaces, dont certains opèrent en temps linéaire par rapport à la taille de l'entrée (c'est le cas des analyseurs par transitions comme MaltParser). Pour ces différentes raisons, les techniques d'analyse syntaxique statistique en dépendances ont été rapidement améliorées et utilisées pour différentes langues, les tâches partagées mentionnées ci-dessus couvrant déjà plusieurs langues typologiquement variées (anglais, arabe, basque, catalan, chinois, grec, hongrois, italien, turc, tchèque). Mais les résultats de ces tâches partagées ont montré que, cette fois-ci pour l'analyse en dépendances, les meilleurs résultats étaient atteints sur les langues à la morphologie flexionnelle la moins riche, et notamment l'anglais.

A.10.2 La communauté SPMRL : analyse syntaxique statistique et informations morphologiques

C'est lors d'une table ronde organisée à l'occasion de la conférence IWPT 2009 que les problématiques spécifiques liées à l'analyse syntaxique de langues à morphologie plus riche que l'anglais ont été discutées spécifiquement. Les travaux présentés à cette occasion sur l'hébreu, l'arabe, le français et l'allemand ont clairement montré qu'il y avait de nombreux points communs entre les difficultés rencontrées par les chercheurs travaillant sur l'analyse syntaxique de langues autre que l'anglais : (i) une couverture lexicale limitée, en raison de la richesse de la morphologie flexionnelle, (ii) une couverture syntaxique limitée, en raison de l'ordre des mots moins fixe, et (iii) plus généralement, des problèmes de dispersion des données plus importants en raison notamment du manque de ressources de grande ampleur.

Cette table ronde a servi de levier à la création d'une communauté active de chercheurs travaillant sur l'analyse syntaxique statistique de langues à morphologie plus riche que l'anglais : c'est la communauté SPMRL (Statistical Parsing for Morphologically Rich

Languages), animée notamment par Djamé Seddah (Alpage, Université Paris-Sorbonne), avec Reut Tsarfaty (Weizmann Institute of Science, Israël) et Sandra Kübler (Indiana University, États-Unis), qui a donné lieu depuis à des workshops annuels organisés autour de conférences internationales majeures du traitement automatique des langues (Seddah *et al.*, 2013b) et à deux tâches partagées. Ces workshops ont été l'occasion de publier des résultats faisant avancer la compréhension des enjeux propres aux langues autres que l'anglais, c'est-à-dire en un sens propres à l'analyse syntaxique des langues en général, et en particulier les différentes façons possibles d'intégrer des informations morphologiques dans le processus d'analyse syntaxique (Goldberg et Tsarfaty, 2008 ; Candito *et al.*, 2010 ; Seddah *et al.*, 2010a ; cf. aussi Green *et al.*, 2013 ; Goldberg et Elhadad, 2013) et de traiter les aspects non projectifs (ou non configurationnels) (Kallmeyer et Maier, 2013). Une autre difficulté rencontrée dans ce contexte est le bas niveau d'adéquation entre mots typographiques (tokens) et mots syntaxiques (feuilles d'un arbre de constituance ou nœuds d'une structure de dépendances ; on lit parfois le terme anglais *tree token*, par opposition aux *source tokens*, nos mots typographiques ou tokens). En effet, si pour l'anglais l'approximation consistant à considérer les mots typographiques comme des mots syntaxiques pour la tâche d'analyse syntaxique fonctionne bien, il n'en est pas de même pour la plupart des autres langues, et notamment dès que les amalgames et les composés sont à la fois nombreux et leur identification non déterministe.

A.10.3 Analyseurs syntaxiques du français : état de l'art début 2014

Les travaux décrits dans ce document, et notamment au chapitre 9, ne se limitent pas à l'analyse syntaxique du seul français, puisque j'ai été amené à travailler également sur l'anglais, l'italien et l'espagnol, puis sur plusieurs dizaines de langues dans le cadre de la campagne d'évaluation UD 2017 (cf. sections 8.2 et 9.5). Toutefois, c'est le français qui a reçu la majorité de notre attention dans ce domaine. Il est donc utile de donner ici un aperçu partiel des analyseurs syntaxiques académiques à la date où nous avons mené plusieurs expériences sur le sujet, notamment celles décrites à la section 9.1. Comme au chapitre 8, nous désignons respectivement par FTB-TRAIN, FTB-DEV et FTB-TEST les sections d'entraînement, de développement et de test du Corpus Arboré de Paris 7 (ou French TreeBank ; Abeillé *et al.*, 2003).

- Analyseurs statistiques entraînés sur le FTB-TRAIN :
  - les analyseurs résultant de l'adaptation au français des l'analyseurs en dépendances MaltParser (Nivre *et al.*, 2007b) et MSTParser de (McDonald *et al.*, 2005) et de l'analyseur en constituants de Berkeley (Petrov *et al.*, 2006 ; Petrov et Klein, 2007), adaptation réalisée dans le cadre de l'initiative Bonsai [48] (Crabbé et Candito, 2008 ; Candito *et al.*, 2009a ; Seddah *et al.*, 2009 ; Candito

48. http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html

et al., 2009b, 2010) ; de nombreux travaux ultérieurs ont permis d'améliorer ces analyseurs syntaxiques (Candito et Seddah, 2010 ; Henestroza Anguiano et Candito, 2011 ; Candito *et al.*, 2011) ;

- l'analyseur en constituants LORG de Attia *et al.* (2010) (analyseur de type Berkeley) ;
- l'analyseur en constituants HyParse, analyseur par transitions développé par Crabbé (2014) ;
- l'adaptation au français l'analyseur en dépendances MATE de Bohnet (2010) (analyseur de type MSTParser) ;
- l'analyseur en dépendances Talismane (Urieli et Tanguy, 2013).
- l'analyseur en dépendances DyALog-SR, analyseur en dépendances de type MaltParser (Villemonte de La Clergerie, 2014).

— Analyseurs symboliques ou hybrides

- Fips, analyseur multilingue et multifonction symbolique développé à l'Université de Genève (Wehrli, 2007), qui repose sur un modèle linguistique inspiré de la grammaire générative (modèle minimaliste), tout en empruntant certaines notions à d'autres formalismes ; il repose sur une base de données lexicales propres ;
- l'analyseur stochastique profond du LPL (Rauzy et Blache, 2009) qui repose sur un modèle de patrons textuels probabilisés ; il repose sur la base lexicale DicoLPL Vanrullen *et al.* (2005) ; Rauzy et Blache (2007) ;
- LEOPAR (Guillaume et Perrier, 2010b), analyseur syntaxique symbolique reposant sur le formalisme des Grammaires d'Interaction (Guillaume et Perrier, 2010a) ; LEOPAR est alimenté par une méta-grammaire permettant la production automatique d'une grammaire d'interaction et par un lexique obtenu par rassemblement de ressources existantes ; LEOPAR ne dispose pas de techniques de désambiguïsation ;
- FRMG (Thomasset et Villemonte de La Clergerie, 2005 ; Villemonte de La Clergerie, 2013), sur lequel nous allons revenir plus en détail, et dont l'utilisation comme guide pour DyALog-SR (cité ci-dessus) donne le système hybride FRMG + DyALog-SR Villemonte de La Clergerie (2013) ;

Le niveau de performance de ces analyseurs n'est pas toujours facile à déterminer. Les campagnes EASy puis PASSAGE d'évaluation des analyseurs syntaxiques (cf. section 9.1.1) ont pu, en leur temps, fournir un aperçu des performances des analyseurs. Mais le format utilisé par ces campagnes ne correspondait que rarement au format et même au type d'analyse produit par les analyseurs des participants — nous avons déjà illustré à la

section 5.3 le fait qu'un désaccord entre les choix linguistiques sous-jacents à un analyseur et ceux sur lesquels reposent un corpus d'évaluation peut conduire à ce qu'une amélioration d'un analyseur conduise à des scores dégradés. On voit donc que les résultats de telles évaluations doivent être nuancées non seulement par la présence inévitable d'erreurs dans le corpus de référence utilisé, mais aussi et surtout par le caractère inévitable de l'introduction d'erreurs ou d'approximations au sein du processus de réinterprétation et conversion post-analyse. Il s'agit en effet d'un processus complexe qui est nécessaire pour passer des représentations natives produites par les analyseurs (en termes de choix linguistiques, de représentations syntaxiques, de jeux d'étiquettes, voire de segmentation) vers le format commun, ici celui des campagnes EASy/PASSAGE — sans compter que cette conversion inclut bien souvent un appauvrissement.

Depuis l'essor des approches statistiques à l'analyse syntaxique du français, les métriques standard sur le plan international ont pu commencer à être utilisées, utilisant FTB-TEST comme corpus d'évaluation. Mais les mêmes difficultés décrites ci-dessus se retrouvent. De plus, l'utilisation du FTB sous la forme d'un corpus annoté en dépendances nécessite sa conversion à partir de son format d'origine, celui d'un corpus annoté en constituants (Candito *et al.*, 2009a), transformation qui est nécessairement imparfaite. Malgré toutes ces difficultés, les scores standard restent des références utiles. Les métriques les plus répandues sont les suivantes : le score d'attachement étiqueté (*Labeled Attachment Score*, LAS), qui est le rapport entre le nombre de dépendances correctes et correctement étiquetées et le nombre total de dépendances (nombre de mots dans la phrase moins un ; les ponctuations sont ignorées), et le score d'attachement non-étiqueté (*Unlabeled Attachment Score*, UAS), calculé comme le LAS mais en ne prenant pas en compte les étiquettes des dépendances. Pour l'analyse en constituants, on a souvent recours à la métrique Parseval.

Un récapitulatif des scores LAS publiés sur le FTB (entraînement sur FTB-TRAIN et évaluation sur FTB-TEST, version FTB-UC) est indiqué à la table A.4 [49]. Comme rappelé à la section 9.1.1.2, la quasi totalité de ces analyseurs syntaxiques utilisent le L*efff*, soit au sein de MElt, soit en exploitant les informations morphologiques et parfois syntaxiques. On constate qu'un analyseur symbolique comme FRMG obtient des performances tout à fait honorables. De plus, Villemonte de La Clergerie (2013, 2014) montre que les performances de FRMG se dégradent moins que celles de l'analyseur de Berkeley lorsque l'on passe d'un corpus journalistique à un corpus médical. Il est donc très intéressant de constater que, contrairement à ce qui a pu être estimé par moments, les techniques d'analyse syntaxique reposant sur des grammaires génératives et des algorithmes performants ont toujours toute leur place, notamment lorsqu'elles sont couplées à des techniques de désambiguïsation statistiques voire neuronales. Il est important de rappeler qu'un

49. Cf. également la table 5a dans Villemonte de La Clergerie (2014).

analyseur statistique ou neuronal ne peut être construit qu'à partir d'un corpus arboré, corpus dont le développement a un coût considérable. Le coût de développement des ressources linguistiques que sont une grammaire comme FRMG et un lexique comme le Le*fff* ne sauraient donc être comparées au seul temps d'entraînement d'un analyseur syntaxique statistique ou neuronal, mais doivent l'être au temps de développement du corpus arboré sous-jacent. Naturellement, les analyseurs statistiques et neuronaux ont des performances et une robustesse indéniables. De plus, l'annotation de corpus arborés peut être confiée à des annotateurs moins pointus dans leur expertise que s'il leur fallait développer une métagrammaire comme celle de FRMG. C'est donc dans l'hybridation entre techniques symboliques, statistiques et désormais neuronales que réside certainement la voie la plus prometteuse. C'est du reste dans cet ordre d'idées que FRMG a été couplé par Villemonte de La Clergerie (2014) avec un analyseur statistique à base de transitions développé à cette fin, DyALog-SR, conduisant, toujours avec le Le*fff* comme ressource lexicale, aux meilleurs scores LAS jamais publiés sur l'analyse syntaxique du français sur le FTB, à savoir 90,25%.

Analyseur	**Références**	**LAS (%)**
Berkeley	Petrov *et al.* (2006) ; Petrov et Klein (2007) ; Candito *et al.* (2010)	86,80
FRMG	Villemonte de La Clergerie (2013) ; cf. section 9.1	87,17
MaltParser	Nivre *et al.* (2007b) ; Candito *et al.* (2010)	87,30
MSTParser	McDonald *et al.* (2005) ; Candito *et al.* (2010)	88,20
Talismane	Urieli et Tanguy (2013)	88,50
LORG + reranking	Le Roux *et al.* (2012b)	89,00
MElt +DyALog-SR	Villemonte de La Clergerie (2014)	89,01
MElt + LORG + reranking	Le Roux *et al.* (2012b)	89,20
MElt + MATE	Bohnet (2010) ; Le Roux *et al.* (2012b)	89,20
(MElt + DyALog-SR) + FRMG	Villemonte de La Clergerie (2014)	90,25

TABLEAU A.4 – Scores LAS de différents analyseurs syntaxiques du français sur la version en dépendances du FTB (entraînement sur FTB-TRAIN et évaluation sur FTB-TEST)

Annexe B

Alexina$_{\text{PARSLI}}$

Cette annexe décrit plus en détails le formalisme Alexina$_{\text{PARSLI}}$, formalisme d'implémentation du modèle formel PARSLI de la morphologie flexionnelle (cf. section 2.2.1 ; Sagot et Walther, 2013). La définition d'Alexina$_{\text{PARSLI}}$ s'est fortement appuyée sur le formalisme morphologique originel d'Alexina (cf. section 2.1). Alexina$_{\text{PARSLI}}$ a vocation, à terme, à remplacer ce dernier, même si à ce stade les deux formalismes coexistent de façon transparente [1]. Naturellement, certains concepts du formalisme morphologique d'Alexina ont été adaptés, étendus et complétés pour les faire correspondre aux concepts définis par PARSLI, comme nous allons le voir dans cette section. Mais les notions issues de variante de classe flexionnelle, de classe de caractères, de règle morphophonologique et les contraintes associées aux règles de réalisation, toutes notions issues du formalisme morphologique originel d'Alexina, ont été conservées en Alexina$_{\text{PARSLI}}$, et parfois améliorées ou généralisées. Par exemple, la syntaxe des règles morphophonologiques a été améliorée mais également étendue afin de permettre également l'écriture de règles purement phonologiques.

B.1 Opérations morphologiques

Dans le formalisme morphologique Alexina d'origine, les seules opérations disponibles pour exprimer les règles de réalisation étaient la préfixation et la suffixation, les opérations plus complexes devant être simulées au moyen de règles morphophonologiques (cf. section 2.1). Alexina$_{\text{PARSLI}}$ permet quant à lui de modéliser également des opérations non-concaténatives directement. Plus précisément, il est possible de définir explicitement dans une grammaire morphologique toute opération requise par les règles de réalisation, y compris si elles sont non-concaténatives. De telles opérations peuvent en effet être

1. Le processus de compilation d'un lexique intensionnel Alexina en lexique extensionnel, qui inclut le processus de flexion, détecte automatiquement si la grammaire morphologique associée au lexique est une grammaire morphologique Alexina ou Alexina$_{\text{PARSLI}}$.

considérées comme partie intégrante du système morphologique à décrire. La suffixation (notée `append=` ou `suffix=`) et la préfixation (notée `left_append=` ou `prefix=`) restent disponibles directement comme en Alexina, en tant qu'opérations de base. Par ailleurs, pour définir une opération complexe utilisable ensuite dans une règle de réalisation, on dispose en supplément d'un opérateur `insert` pour insérer un segment et d'un opérateur `replace` pour remplacer un segment par un autre. Des exemples d'opérations complexes pour des grammaires Alexina PARSLI du maltais et du latin sont fournies à la figure B.2 et glosées ci-dessous. La définition des opérations morphologiques complexes peut, comme pour les règles morphophonologiques, faire usage des classes de caractères héritées d'Alexina. La figure B.1 présente deux définitions de classes de caractères.

```
<letterclass name="C" letters="b␣ċ␣d␣f␣ġ␣g␣għ␣ħ␣h␣j␣k␣l␣m␣n␣p␣q␣r␣s␣t␣v␣w␣x␣ż␣z␣'␣"/>
<letterclass name="V" letters="a␣e␣i␣o␣u␣ie"/>
```

FIGURE B.1 – Deux classes de lettres définies dans la grammaire morphologique de MaltLex

Expliquons désormais les opérations morphologiques définies à la figure B.2. Sauf information complémentaire (cf. ci-dessous), les classes de caractères utilisées dans l'input (`source=`) et l'output (`target=`) doivent se correspondre, c'est-à-dire se succéder à l'identique. Les caractères entre identifiants de classes sont en revanche traités comme des constantes. Pour remplacer un *a* par un *e* entre deux consonnes, et si la classe de caractères `C` a été définie comme listant les lettres dénotant des consonnes, on peut par exemple utiliser la règle `<replace source="[:C:]a[:C:]" target="[:C:]e[:C:]" />` Mais différents mécanismes permettent d'aller plus loin.

La définition de `deleteV1`, opération définie pour le traitement de certains radicaux verbaux du maltais, stipule par exemple que pour un input donné ayant une structure consonne-voyelle de type CVCVC, appliquer cette opération produit une output de structure CCVC après effacement de la première voyelle : c'est le sens du `0` inséré dans l'identifiant de classe de caractères `[0:V:]`. La règle suivante dans la définition de `deleteV1`, qui s'applique aux inputs de structure CVCV, efface également la première voyelle.

Un identifiant de classe de caractères peut également être numérotée par un entier supérieur ou égal à 1 afin de pouvoir le reprendre explicitement côté output, en ignorant donc le parallélisme de la succession des classes de lettres entre input et output. On peut ainsi par exemple inverser deux consonnes (`<replace source="[1:C:][2:C:]" target="[2:C:][1:C:]" />`) ou dupliquer une consonne initiale (début et fin de mot étant notés #) en insérant un *e* (`<replace source="#[1:C:]" target="#[1:C:]e[1:C:]" />`)

Comme en Alexina, Alexina PARSLI permet également de définir des appariements entre classes de caractères. On peut ainsi définir un appariement `lengthening` entre

une classe de caractères indiquant des voyelles brèves et une classe de caractères indiquant des voyelles longues, en appariant chaque voyelle brève avec la voyelle longue qui lui correspond. Dans ce cas, si le parallélisme entre classes de caractères de l'input et de l'output fait se correspondre une voyelle brève et une voyelle longue, l'appariement `lengthening` est déclenché et procède au remplacement de la voyelle brève correspondante en une voyelle longue. Naturellement, si aucun appariement ne convient, la définition de l'opération est invalide. Par exemple, remplacer entre deux consonnes une séquence voyelle brève + *h* par une voyelle longue donnerait `<replace source="[:C:][:Vshort:]h[:C:]" target="#[C:][:Vlong:][:C:]" />`, pour peu qu'on ait défini les classes `Vshort` et `Vlong` ainsi que leur appariement.

Enfin, une opération peut être définie comme ayant un ou plusieurs arguments. Ces arguments remplacent les symboles « _ » placées côté output, dans l'ordre gauche-droite. Par exemple, l'opération `deleteV1changeV2` de la figure B.2 supprime les deux voyelles d'un input de la forme CVCVC ou CVCV et met à la place de la seconde l'argument avec lequel elle est appelée. Ainsi, si une règle de réalisation fait appel à `deleteV1changeV2(i)`, la première voyelle de l'input sera supprimée et la seconde remplacée par *i*.

La règle `redup-initial`, définie pour le latin comme indiqué à la figure B.2, combine ces différentes possibilités en réduplicant la consonne initiale de son input, insérant son premier argument entre la consonne initiale et sa réduplication, supprimant la voyelle qui la suit et la remplaçant par son deuxième argument. Ainsi, exécuter `redup-initial(e,e)` sur le radical *fall-* du verbe latin FALLO 'décevoir' produit le radical réduplicé *fefell-*.

Lors de l'exécution d'une d'opération, la première règle appliquable fournie par sa définition s'applique, et les règles suivantes sont ignorées, sauf si la règle appliquée dispose d'un attribut `stop="0"`. Si l'on cherche à effectuer une opération à un input pour laquelle aucune des règles ne peut être appliquée, l'opération échoue.

```
<!-- maltais (MaltLex) -->
<operation_definition name="deleteV1">
  <replace source="[:C:][0:V:][:C:][:V:][:C:]" target="[:C:][:C:][:V:][:C:]" />
  <replace source="[:C:][0:V:][:C:][:V:]" target="[:C:][:C:][:V:]" />
</operation_definition>
<operation_definition name="deleteV1changeV2">
  <replace source="[:C:][0:V:][:C:][0:V:][:C:]" target="[:C:][:C:]_[:C:]" />
  <replace source="[:C:][0:V:][:C:][0:V:]" target="[:C:][:C:]_" />
</operation_definition>

<!-- latin (Leffla) -->
<operation_definition name="redup-initial">
  <replace source="#[1:C:][0:V:]" target="#[1:C:]_[1:C:]_"/>
</operation_definition>
```

FIGURE B.2 – Quelques opérations morphologiques en Alexina$_{\text{PARSLI}}$ (données de MaltLex et du Le*ff*la)

On notera que tous les mécanismes disponibles pour exprimer une règle au sein de la définition d'une opération morphologique sont également disponibles pour énoncer des règles morphophonologiques. Alexina$_{\text{PARSLI}}$ permet ainsi d'exprimer des règles morphophonologiques plus complexes que le formalisme morphologique Alexina d'origine. De plus, Alexina$_{\text{PARSLI}}$ permet l'écriture de règles phonologiques, qui n'incluent pas de frontière de morphe (il suffit d'écrire une règle ne comportant pas de signe « _ ») [2].

B.2 Allomorphie radicale, radicaux supplétifs et formes supplétives

Dans le formalisme morphologique Alexina d'origine, chaque entrée lexicale était considérée comme s'appuyant sur un radical unique. Rendre compte de l'*allomorphie radicale* nécessitait de contourner cette limitation en faisant un usage *ad hoc* et linguistiquement immotivé de dispositifs comme les règles morphophonologiques [3]. En Alexina$_{\text{PARSLI}}$, l'allomorphie radicale est traitée directement par deux mécanismes distincts, l'un pour les alternances régulières (comme dans les langues iraniennes) et l'autre pour les alternances irrégulières (comme pour le verbe français ALLER, illustré à la figure B.4) [4].

L'allomorphie régulière est modélisée à l'aide de règles de réalisation au niveau radical dans la grammaire, qui permettent de dériver les radicaux à partir de l'input, soit directement soit à partir d'un autre radical : les règles de construction des radicaux sont ainsi organisés sous forme arborescente, à l'image de ce que proposent (Bonami et Boyé, 2003). La figure B.3 contient deux règles de ce type pour modéliser l'allomorphie radicale régulière de certains verbes du maltais, laquelle est illustrée dans les deux paradigmes de la table B.1. Au sein des paradigmes des verbes maltais du premier binyan, l'allomorphie radicale implique jusqu'à six radicaux distincts (RAD1 à RAD6). Les exemples de la table B. 1 en impliquent quatre, indiqués par quatre couleurs de fond distinctes dans le paradigme (dans ces paradigmes, RAD1 et RAD2 sont syncrétiques, de même que RAD5 et RAD6).

L'autre façon de représenter les alternances radicales en Alexina$_{\text{PARSLI}}$ consiste à les spécifier explicitement dans l'entrée lexicale. C'est la façon naturelle de représenter l'alternance radicale irrégulière, c'est-à-dire les cas où la construction de radicaux n'est

2. On peut également noter que les règles morphophonologiques peuvent être utilisées pour traiter de la notion de morphe sous-spécifié. Par exemple, en turc, le pluriel des noms est exprimé par le suffixe *-lar* ou *-ler*, selon les voyelles du nom : il y a harmonie vocalique. On peut parfaitement imaginer modéliser ceci à l'aide d'un suffixe unique *-lær*, où *æ* est un graphème abstrait (sous-spécifié), couplé à des règles phonologiques réécrivant *æ* en *a* ou *e* selon la dernière voyelle du radical.

3. Par exemple, on peut définir des règles morphophonologiques artificielles qui jouent le rôle de règles de production de radicaux, à partir d'un radical de base. Dans ce cas, chaque règle réalisationnelle peut suffixer au radical de base un indice du type de radical utilisé par la forme, cet indice étant destiné à être interprété par des règles morphophonologiques artificielles qui sont en réalité des règles de réalisation des radicaux. Cette solution temporaire et *ad hoc* a par exemple été utilisée sur le persan dans le lexique PerLex.

4. Nous renvoyons le lecteur à (Walther, 2013b) pour une discussion sur la façon dont on peut distinguer allomorphie régulière et allomorphie irrégulière.

		RASS ‘presser’	MESS ‘toucher’			RASS ‘presser’	MESS ‘toucher’
RAD2	PFV 1.SG	*rasséjt*	*messéjt*	RAD5	IPFV 1.SG	*nróss*	*nmíss*
	PFV 2.SG	*rasséjt*	*messéjt*		IPFV 2.SG	*tróss*	*tmíss*
	PFV 1.PL	*rasséj.na*	*messéj.na*		IPFV 3.M.S	*jróss*	*jmíss*
	PFV 2.PL	*rasséj.tu*	*messéj.tu*		IPFV 3.F.S	*tróss*	*tmíss*
RAD1	PFV 3.M.S	*rás*	*més*		IPFV 1.PL	*nrós.su*	*nmís.su*
RAD3	PFV 3.F.S	*rás.set*	*més.set*	RAD6	IPFV 2.PL	*trós.su*	*tmís.su*
RAD4	PFV 3.PL	*ras.sé:w*	*mes.sé:w*		IPFV 3.PL	*jrós.su*	*jmís.su*
	sous-paradigme perfectif				*sous-paradigme imperfectif*		

TABLEAU B.1 – Paradigmes pour les verbes maltais RASS et MESS

```
<table name="CVCC" rads="[:C:][:V:][:C:][:C:]">
  <item name="S1"/>
  <item name="S2" source="S1" append="ej"/>
  <item name="S3" source="S1" operation="" />
  <item name="S4" source="S1" append="e"/>
  <item name="S5" source="S1" operation="changeV1(o)" rads="[:C:]a[:C:][:C:]"/>
  <item name="S5" source="S1" operation="changeV1(i)" rads="[:C:]e[:C:][:C:]"/>
  <item name="S5" source="S1" operation="changeV1(i)" rads="[:C:]i[:C:][:C:]"/>
  <item name="S6" source="S5" operation="" />
</table>
```

FIGURE B.3 – Allomorphie radicale régulière (données de MaltLex, d’après Camilleri et Walther (2012))

pas couverte par la grammaire mais constitue une irrégularité lexicale de l’entrée considérée : ce sont des radicaux supplétifs. Si l’on considère que les verbes français ont douze radicaux (souvent syncrétiques), suivant ainsi Bonami et Boyé (2003) et en conservant les mêmes identifiants de radicaux allant de S1 à S12, chacun de ces radicaux correspond à un slot après le symbole « / » dans l’entrée lexicale, triés et séparés par une virgule. Par exemple, dans le cas du verbe ALLER, et comme indiqué à la figure B.4, les radicaux supplétifs spécifiés dans le lexiques sont les radicaux S2 *va-*, S7 *aill-* et S10 *i-*. Les autres radicaux sont obtenus à partir de ces trois radicaux ou du radical de la forme de citation, *all-*, par les règles gérant l’allomorphie régulière dans la grammaire. Ainsi, le radical S3 est déduit du radical S2 par syncrétisme.

```
aller   v:23r/,va,,,,,aill,,,i
dire    v:3re/dis,,di,,,,,,,,,dit/2.pl.prs.ind=dites
```

FIGURE B.4 – Allomorphie radicale irrégulière et formes supplétives (données de l’une des versions du Le*fff* utilisées par Sagot et Walther (2011) et Walther et Sagot (2011a))

Le formalisme morphologique d'origine d'Alexina n'avait pas non plus de moyen d'encoder la *supplétion de forme*, c'est-à-dire le fait qu'une forme du paradigme d'un lemme donné ne soit construite par l'application des règles de réalisation présentes dans la grammaire morphologiques, pas même via un radical supplétif. Ainsi, ce phénomène non-canonique devait être modélisé d'une façon *ad hoc*[5]. En Alexina𐑐𐑸𐑮𐑕𐑤, il est possible de lister explicitement les formes supplétives dans l'entrée lexicale, après le deuxième symbole « `/` ». Elles remplacent alors toute forme produites par la grammaire pour les cases concernées du paradigme. La figure B.4 illustre ceci sur l'exemple du verbe DIRE, qui a la forme irrégulière *dites* pour la case 2.PL.PRS.IND à la place de la forme régulière *disez*.

Dans le cas de formes supplétives surabondantes, le mécanisme déjà disponible dans le formalisme Alexina d'origine est préservé : elles peuvent être listées explicitement comme telles. Ainsi, pour le pluriel de l'adjectif MARRON, et en plus de la forme régulière *marrons*, le L*efff* liste à part la forme supplémentaire *marron*. La surabondance régulière est quant à elle traitée au moyen de sous-schèmes flexionnels multiples (cf. ci-dessous).

B.3 Niveaux réalisationnels, zones flexionnelles et schèmes flexionnels

Une autre spécificité d'Alexina𐑐𐑸𐑮𐑕𐑤 par rapport au formalisme d'origine d'Alexina est qu'il permet de représenter le processus de construction des formes au moyen de plusieurs niveaux réalisationnels (`level`), comme le propose le modèle 𐑐𐑸𐑮𐑕𐑤, faisant suite en cela, quoique pas de la même façon, à des travaux tels que ceux de Kiparsky (1982), Anderson (1992) ou Stump (2001). Une grammaire morphologique Alexina𐑐𐑸𐑮𐑕𐑤 doit ainsi définir au plus un niveau radical (`type="stem"`) et un nombre quelconque, y compris nul, de niveaux thématiques (`type="theme"`) et de niveaux d'exponence (`type="exponent"`).

𐑐𐑸𐑮𐑕𐑤 définit les espaces partitionnants comme des sous-ensembles des structures de traits morphosyntaxiques canoniquement réalisées par les lemmes d'une catégorie donnée. Ils constituent l'un des moyens par lesquels on peut faire référence à des zones réalisationnelles telles que définies par 𐑐𐑸𐑮𐑕𐑤. En Alexina𐑐𐑸𐑮𐑕𐑤, ces espaces partitionnants peuvent être définis pour chaque niveau. Nous illustrons ce mécanisme à la figure B.5 sur le maltais au niveau radical et sur le latin au niveau d'exponence.

L'une des innovations majeures de 𐑐𐑸𐑮𐑕𐑤, et par voie de conséquence d'Alexina𐑐𐑸𐑮𐑕𐑤, est l'introduction des *zones réalisationnelles*. En Alexina𐑐𐑸𐑮𐑕𐑤, elles peuvent être définies de deux façons différentes. La première consiste à définir directement des zones réalisationnelles (`zone`) au sein d'un niveau réalisationnel donné. La seconde consiste à

5. Soit en assignant à l'entrée une classe flexionnelle qui ne générerait pas toutes les formes du paradigmes, voire aucune d'entre elles, et en inventoriant à part les formes supplétives, soit en assignant à l'entrée une classe flexionnelle *ad hoc* qui considère la quasi-totalité voire la totalité des formes comme étant un exposant (suffixal) qui se combine avec un radical presque vide, voire vide.

```
<!-- maltais (MaltLex) -->
<level type="stem" level="1">
    <partitionspace name="S1" features="3.m.sg.pfv"/>
    <partitionspace name="S2" features="1.pfv|2.pfv"/>
    <partitionspace name="S3" features="3.f.sg.pfv"/>
    <partitionspace name="S4" features="3.pl.pfv"/>
    <partitionspace name="S5" features="sg.ipfv"/>
    <partitionspace name="S6" features="pl.ipfv"/>

<!-- latin (Leffla) -->
<level type="exponent" level="3">
   <partitionspace name="I1" features="ipfv.ind|ipfv.sbjv|prs.inf"/>
   <partitionspace name="I2" features="pfv.ind|pfv.sbjv|pst.inf"/>
   <partitionspace name="I3" features="prs.ptcp|fut.ptcp|fut.inf|sup|pst.ptcp|grv|grd"/>
```

FIGURE B.5 – Définition d'espaces partitionnants (données de MaltLex et du Le*ff*la)

faire usage de classes réalisationnelles, qui sont alors définies comme des assemblages cohérents de zones réalisationnelles d'un niveau donné qui sont utilisées ensemble par un nombre significatif d'entrées lexicales : il s'agit d'une notion dérivée, qui s'apparente à la notion classique de classe flexionnelle [6]. Les zones réalisationnelles (`zone`) impliquées dans une classe réalisationnelle sont alors spécifiées à l'intérieur de la définition de la classe (`table`).

Dans le formalisme morphologique Alexina d'origine, les entrées intentionnelles sont associées à des classes flexionnelles. Alexina*ParSLi* implémente quant à lui le point de vue de *ParSLi*, selon lequel une entrée lexicale est associée à des zones réalisationnelles à travers la notion de *schème réalisationnel*, comme expliqué plus haut. Un schème réalisationnel est composé d'au moins un *sous-schème*, chaque sous-schème étant défini avant tout comme une liste de spécifications de zones réalisationnelles (`realzone`), une par niveau réalisationnel pertinent (par exemple, une zone de niveau radical et une zone d'exponence). On peut spécifier une zone réalisationnelle de deux façons différentes : soit directement, par son nom, soit par un couple formé du nom d'une zone ou d'une classe réalisationnelle et d'un espace partitionnant : la zone ainsi dénotée est la projection de la zone ou de la classe sur l'ensemble des cases constituant l'espace partitionnant [7]. Cette notion est illustrée à la figure B.6). On notera que ces `realzones`, en Alexina*ParSLi*, peuvent également être munies contraintes sur leur input [8] et d'identifiants de « variantes » à spécifier dans le lexique pour contrôler leur applicabilité. Ainsi, un sous-schème peut invoquer plusieurs `realzones` d'un même niveau, munies de contraintes différentes ; pour un niveau donné, la première `realzones` appliquable est utilisée. Il s'agit là de l'un

6. La notion de classe flexionnelle peut être définie comme étant identique à celle de classe réalisationnelle de niveau exponence, à condition de négliger les éventuelles alternances radicales.

7. On peut également invoquer directement une classe réalisationnelle sans espace partitionnant. Elle est alors considérée dans son ensemble comme une zone, même si elle contient elle-même plusieurs zones.

8. Contraintes positives indiquées par `rads=`, contraintes d'exclusion indiquées par `rads_except`.

des dispositifs de factorisation mentionnés précédemment. Chaque sous-schème peut produire au plus une forme pour une structure de traits morphosyntaxiques donnée, c'est-à-dire pour une case donnée. La surabondance régulière requiert donc des schèmes qui contiennent plusieurs sous-schèmes. De plus, chaque schème se voit assigné une *catégorie morphologique*. Cette dernière permet de connaître, par un moyen décrit ci-dessous, l'inventaire des structures de traits morphosyntaxiques réalisées par les lemmes associés à ce schème, c'est-à-dire l'ensemble des cases de leur paradigme.

La figure B.6 illustre la façon dont l'un des schèmes les plus fréquents est défini dans la grammaire morphologique du latin utilisée par Le*ff*la et dans l'une des versions Alexina PARSLI du Le*fff*. Le schème issu du Le*ff*la fait tout d'abord usage de la classe réalisationnelle `REG-EQ-T` au niveau radical et de la classe réalisationnelle thématique des verbes à thème en *a*. Il utilise alors à un premier niveau d'exponence la classe réalisationnelle complète des exposants de l'actif. Un dernier niveau d'exponence s'applique alors, mais uniquement pour les cases participiales, en utilisant des règles réalisationnelles adjectivales. La mention `on_failure="skip"` indique que, contrairement au comportement standard, un échec d'application d'une `realzone` ne doit pas être considérée comme indiquant l'échec de la construction de la forme : ceci permet de ne prendre en compte le deuxième niveau d'exponence que pour les formes participiales, les seules pour lesquelles il est pertinent.

Le schème issu du Le*fff* reprend quant à lui l'exemple des verbes en *–ayer* comme BALAYER. On retrouve les trois sous-schèmes de la figure 2.3, qui permettent de modéliser la surabondance de cette classe de verbes sur une partie de leur paradigme (les deux derniers sous-schèmes produisent des formes pour les mêmes cases, celles de l'espace partitionnant de la zone d'exponence *v1,2*).

Comme nous l'avons vu précédemment, PARSLI modélise explicitement le phénomène de décalage, dont le pluriel du serbo-croate BRAT 'frère' a été donnée comme illustration, au moyens de fonctions de transfert. Chaque `realzone` est ainsi associée à une fonction de transfert qui prend en entrée la structure de traits morphosyntaxiques à réaliser et fournit en sortie la structure de traits, que l'on peut qualifier de morphologiques, à fournir aux sous-schèmes pour que soient sélectionnées les règles de réalisation à appliquer. Canoniquement, cette fonction de transfert est la fonction identité. Elle n'est dans ce cas pas indiquée. Dans les autres cas, elle doit être invoquée au moyen de l'attribut `transfer` associé à la `realzone` concernée, la valeur de cet attribut étant le nom d'une fonction de transfert définie par ailleurs. La définition de la fonction de transfert consiste en une liste de couples de structures de traits, l'une morphosyntaxique (`source`) et l'autre morphologique (`target`) permettant d'obtenir les traits à fournir au sous-schème à partir de ceux à réaliser. La figure B.7 illustre la façon dont on peut définir la fonction de transfert à l'œuvre dans le sous-schème utilisé par le nom serbo-croate BRAT 'frère' pour

```
<!-- latin (Leffla) -->
 <pattern name="v-aA-REG-EQ-T" cat="v">
  <subpattern>
   <realzone level="1" table="REG-EQ-T"/>
   <realzone level="2" table="a"/>
   <realzone level="3" table="v-A" />
   <realzone level="4" table="adj-oa" features="ptcp.fut|grv|grd|ptcp.pst" on_failure="skip
       " />
   <realzone level="4" table="adj-c" features="ptcp.prs" on_failure="skip"/>
  </subpattern>
 </pattern>

<!-- Français (Lefff modifié) -->
 <pattern name="v-aA-REG-EQ-T" cat="v">
  <subpattern>
   <realzone level="1" table="v-stem-std"/>
   <realzone level="2" table="v1,1"/>
  </subpattern>
  <subpattern>
   <realzone level="1" table="v-stem-std"/>
   <realzone level="2" table="v1,2"/>
  </subpattern>
  <subpattern>
   <realzone level="1" table="v-stem-y-becomes-i"/>
   <realzone level="2" table="v1,2"/>
  </subpattern>
 </pattern>
```

FIGURE B.6 – Exemples de définitions de schèmes (données de Le*ff*la et d'une version Alexina ParSLI du Le*fff*)

la réalisation de ses formes du pluriel, qui associe le trait à réaliser PL au trait SG, ce nom utilisant pour réaliser ses formes du pluriel des règles de réalisation associées à des structures de traits du singulier. On notera que ces fonctions de transfert peuvent également permettre, comme étudié par Walther *et al.* (2013), de modéliser un processus flexionnel à l'aide de traits morphomiques, c'est-à-dire de traits morphologique qui n'ont pas de correspondance naturelle avec les traits morphosyntaxiques exprimés, mais qui permettent de capturer des généralisations pertinentes et de simplifier la grammaire morphologique.

```
 <transfer name="transfer-vi-t">
  <transfer_rule source="pl" target="sg"/>
 </transfer>
```

FIGURE B.7 – Exemple de fonction de transfert

B.4 Traits morphosyntaxiques et définition des cases des paradigmes

Dans le formalisme morphologique Alexina d'origine, les traits morphosyntaxiques n'apparaissaient que sous forme d'étiquettes associées aux règles de réalisation (qui n'étaient que des règles d'exponence). En PARSLI, modèle réalisationnel, chaque forme est considérée comme étant la réalisation d'une structure de traits morphosyntaxiques. AlexinaPARSLI modélise donc explicitement ces structures de traits. L'inventaire des cases propre à une catégorie morphologique est calculé via un mécanisme d'unification. Pour une catégorie morphologique donnée (`category`), les cases à réaliser sont obtenues par la combinaison de couples attribut-valeur mutuellement compatibles. Ainsi, une grammaire morphologique AlexinaPARSLI spécifie pour chaque catégorie l'inventaire des attributs pertinents, la liste des valeurs possibles pour chacun d'eux, ainsi que des règles d'exclusion, qui peuvent être de trois type : une valeur pour un certain attribut peut être déclarée non compatible avec une valeur pour un autre attribut, une valeur pour un certain attribut peut être déclarée non compatible avec un attribut, et une structure de trait complète peut être déclarée invalide.

Enfin, on peut définir des ensembles de structures de traits morphosyntaxiques et les associer à des entrées lexicales, modélisant ainsi la déficience. Par exemple, en français, les verbes impersonnels peuvent être associés à un ensemble de traits `impers`, qui ne contient que des cases du paradigme verbal dont la structure de traits morphologiques s'unifient avec 3.SG.

B.5 Règles de réalisation

Nous avons décrit ci-dessus la façon dont on peut définir en AlexinaPARSLI des opérations morphologiques, et comment elles peuvent être invoquées, y compris par les règles de réalisation. Contrairement à ce qui se passe dans le formalisme morphologique Alexina, AlexinaPARSLI associe les règles de réalisation à des structures de traits morphologiques. À un niveau donné, une règle de réalisation sera appliquée si sa structure de traits s'unifie avec la structure de traits de la forme à réaliser.

En AlexinaPARSLI, il est possible, comme en Paradigm Function Morphology, de définir des blocs de règles au sein d'une zone ou d'une classe. Dans chaque bloc, une et une seule règle sera appliquée pour la réalisation d'une structure de traits donnée. Dans l'exemple maltais de la figure B.8, qui illustre l'unique classe de niveau exponence du lexique MaltLex, nous utilisons deux blocs de règles (`block="1"`) and (`block="2"`). Le premier bloc réalise l'aspect et la personne, alors que le second réalise le nombre (suffixe *-u* pour les formes du pluriel). Comme en Paradigm Function Morphology, AlexinaPARSLI permet l'écriture de règles *porte-manteau*, règles prioritaires qui s'étendent sur plusieurs blocs consécutifs. Dans cet exemple, `block="1-2"` permet à une règle de court-circuter

les deux blocs. Enfin, si, au sein d'un bloc, plusieurs règles pourraient s'appliquer pour la production d'une même forme, la première s'applique (contrairement à ce qui se passe en Paradigm Function Morphology, où est appliquée une priorité paninienne, c'est-à-dire que la plus spécifique s'applique).

```
<level type="exponent" level="3">
  <table name="exponence" rads="">
    <item block="1-2" suffix="na" features="1.pl.pfv"/>
    <item block="1-2" suffix="et" features="3.f.sg.pfv"/>
    <item block="1" suffix="t" features="1.sg.pfv|2.pfv"/>
    <item block="1" prefix="n" features="1.ipfv"/>
    <item block="1" prefix="t" features="2.ipfv"/>
    <item block="1" prefix="t" features="3.f.sg.ipfv"/>
    <item block="1" prefix="j" features="3.ipfv"/>
    <item block="2" suffix="u" features="pl"/>
  </table>
</level>
```

FIGURE B.8 – Règles de réalisation : l'unique table d'exponence de MaltLex

Bibliographie

Abeillé, A. 2002, *Une grammaire électronique du français*, CNRS Editions, Paris, France.

Abeillé, A., L. Clément et F. Toussenel. 2003, « Building a treebank for French », dans *Treebanks*, édité par A. Abeillé, Kluwer Academic Publishers, Dordrecht, Pays-Bas.

Ackerman, F., J. P. Blevins et R. Malouf. 2009, « Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter », dans *Analogy in Grammar : Form and Acquisition*, édité par J. P. Blevins et J. Blevins, Oxford University Press, Oxford, Royaume-Uni, p. 54–82.

Ackerman, F. et R. Malouf. 2013, « Morphological organization: The low conditional entropy conjecture », *Language*, vol. 89, p. 429–464.

Adda, G., B. Sagot, K. Fort et J. Mariani. 2011, « Crowdsourcing for Language Resource Development : Critical Analysis of Amazon Mechanical Turk Overpowering Use », dans *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, Poznań, Pologne.

Adolphs, P. 2008, « Acquiring a poor man's inflectional lexicon for German », dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Agić, Z., K. Šojat et M. Tadić. 2010, « An experiment in verb valency frame extraction from croatian dependency treebank », dans *Proceedings of the 32nd International Conference on Information Technology Interfaces (ITI)*, p. 55–60.

Aho, A., M. Lam, R. Sethi et J. Ullman. 2007, *Compilateurs : principes, techniques et outils (2ème édition)*, Pearson Education. Traduit en français par P. Deschamp, B. Lorho, B. Sagot et F. Thomasset.

Aho, A. V., R. Sethi et J. Ullman. 1986, *Compilers, principles, techniques, and tools*, Addison-Wesley, Reading, Massachusetts, États-Unis.

Aho, A. V. et J. Ullman. 1972, *The Theory of Parsing, Translation and Compiling (vol. 1)*, Prentice Hall, Englewood Cliffs, New Jersey, États-Unis.

Al-Rfou, R., B. Perozzi et S. Skiena. 2013, « Polyglot : Distributed Word Representations for Multilingual NLP », dans *Proceedings of the 7th Conference on Computational Natural Language Learning (CoNLL 2013)*, Sofia, Bulgarie, p. 183–192.

Albright, A. 2008, « Inflectional paradigms have bases too : Arguments from Yiddish », dans *Inflectional identity*, édité par A. Bachrach et A. Nevins, Oxford University Press, p. 271–312.

Álvarez, C., P. Alvariño, A. Gil, T. Romero, M. P. Santalla et S. Sotelo. 1998, « AVALON, una gramática formal basada en corpus », dans *Procesamiento del Lenguaje Natural (Actas del XIV CONGRESO de la SEPLN)*, Alicante, Espagne, p. 132–139.

Anderson, S. R. 1992, *A-morphous Morphology*, Cambridge University Press, Cambridge, Royaume-Uni.

Apidianaki, M. et B. Sagot. 2012, « Applying cross-lingual WSD to wordnet development », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Apidianaki, M. et B. Sagot. 2014, « Data-driven synset induction and disambiguation for wordnet development », *Language Resources and Evaluation*, vol. 48, nº 4, p. 655–677.

Aronoff, M. 1994, *Morphology by Itself: Stems and Inflectional Classes*, Linguistic Inquiry Monographs, MIT Press.

Aronoff, M. et K. Fudeman. 2005, *What is morphology?*, Blackwell, Oxford, Royaume-Uni.

Arun, A. et F. Keller. 2005, « Lexicalization in crosslinguistic probabilistic parsing: The case of french », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, p. 306–313.

Attia, M., J. Foster, D. Hogan, J. L. Roux, L. Tounsi et J. van Genabith. 2010, « Handling unknown words in statistical latent-variable parsing models for arabic, english and french », dans *Proceedings of the NAACL/HLT 2010 Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, Californie, États-Unis.

Baayen, R. 2009, « The morphological complexity of simplex nouns », *Linguistics*, vol. 35, nº 5, p. 861–878.

Baayen, R. H., R. Piepenbrock et L. Gulikers. 1993, « The CELEX lexical data base on CD-ROM », .

Baerman, M. 2007, « Morphological Typology of Deponency », dans *Deponency and Morphological Mismatches*, vol. 145, édité par M. Baerman, G. G. Corbett, D. Brown et A. Hippisley, The British Academy, Oxford University Press, Oxford, Royaume-Uni, p. 1–19.

Baerman, M., D. Brown et G. G. Corbett. 2015, « Understanding and measuring morphological complexity: An introduction », dans *Understanding and Measuring Morphological Complexity*, édité par M. Baerman, D. Brown et G. G. Corbett, Oxford University Press, Oxford, Royaume-Uni.

Bailly, C. 1944, *Linguistique Générale et linguistique française (2ème édition)*, Francke, Berne, Suisse.

Baker, C. F., C. J. Fillmore et J. B. Lowe. 1998, « The Berkeley FrameNet project », dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, Montréal, Québec, Canada, p. 86–90.

Ballesteros, M., C. Dyer et N. A. Smith. 2015, « Improved Transition-based Parsing by Modeling Characters instead of Words with LSTMs », dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisbonne, Portugal, p. 349–359.

Balvet, A., L. Barque, M.-H. Condette, P. Haas, R. Huyghe, R. Marín et A. Merlo. 2011, « Nomage: an electronic lexicon of french deverbal nouns based on a semantically annotated corpus », dans *Proceedings of the International Workshop on Lexical Recources (WoLeR 2011)*, Ljubljana, Slovénie, p. 8–15.

Bane, M. 2008, « Quantifying and measuring morphological complexity », dans *Proceedings of the 26th West Coast Conference on Formal Linguistics (WCCFL 2008)*, Berkeley, Californie, États-Unis, p. 69–76.

Bangalore, S., P. Boullier, A. Nasr, O. Rambow et B. Sagot. 2009, « MICA : A Probabilistic Dependency Parser Based on Tree Insertion Grammars », dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*, Boulder, Colorado, États-Unis.

Baranes, M. 2012, « Vers la correction automatique de textes bruités : Architecture générale et détermination de la langue d'un mot inconnu », dans *Actes de la 19ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, Grenoble, France.

Baranes, M. 2015, *Normalisation orthographique de corpus bruités*, Thèse de doctorat, Université Denis–Diderot Paris 7, Paris, France.

Baranes, M. et B. Sagot. 2014a, « A language-independent approach to extracting derivational relations from an inflectional lexicon », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Baranes, M. et B. Sagot. 2014b, « Normalisation de textes par analogie : le cas des mots inconnus », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France, p. 137–148.

Bartens, H.-H. 1989, *Lehrbuch der Saamischen (Lappischen) Sprache*, Helmut Buske, Hamburg, République Fédérale d'Allemagne.

Barthélemy, F., P. Boullier, P. Deschamp et É. de la Clergerie. 2001, « Guided parsing of range concatenation languages », dans *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*, Toulouse, France, p. 42–49.

Baudouin de Courtenay, J. N. I. 1895, *Versuch einer Theorie phonetischer Alternationen: Ein Kapitel aus der Psychophonetik*, Trübner, Strasbourg, Empire allemand.

Beaufort, R., S. Roekhaut, L.-A. Cougnon et C. Fairon. 2010, « A Hybrid Rule/Model-Based Finite-State Framework for Normalizing SMS Messages », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède, p. 770–779.

Béchet, D. et A. Foret. 2009, « PPQ: a pregroup parser using majority composition », dans *Proceedings of the ESSLLI 2009 Workshop on Parsing with Categorial Grammars*, édité par Timothy Fowler et Gerald Penn, Bordeaux, France, p. 33–37.

Béchet, F., B. Sagot et R. Stern. 2011, « Coopération de méthodes statistiques et symboliques pour l'adaptation non-supervisée d'un système d'étiquetage en entités nommées », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France.

Beekes, R. 2009, *Etymological Dictionary of Greek*, Brill.

Beesley, K. R. et L. Karttunen. 2003, *Finite State Morphology*, Studies in Computational Linguistics, CSLI Publications.

Beinborn, L., T. Zesch et I. Gurevych. 2013, « Cognate production using character-based machine translation », dans *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japon.

Bengio, Y., R. Ducharme, P. Vincent et C. Janvin. 2003, « A neural probabilistic language model », *Journal of Machine Learning Research*, vol. 3, nº 1, p. 1137–1155.

Beniamine, S., O. Bonami et B. Sagot. 2015, « Information-theoretic inflectional classification », dans *Présentation orale au 1er Quantitative Morphology Meeting (QMM1)*, Belgrade, Serbie.

Beniamine, S., O. Bonami et B. Sagot. 2018, « Inferring inflection classes with description length », *Journal of Language Modelling*, vol. 5, nº 3.

Beniamine, S. et B. Sagot. 2015, « Segmentation strategies for inflection class inference », dans *Décembrettes 9, Colloque international de morphologie*, Toulouse, France.

Benzitoun, C., A. Dister, K. Gerdes, S. Kahane, P. Pietrandrea et F. Sabio. 2010, « tu veux couper là faut dire pourquoi — propositions pour une segmentation syntaxique du français parlé », dans *Actes du 2ème Congrès Mondial de Linguistique Française (CMLF 2010)*, La Nouvelle-Orléans, Louisiane, États-Unis.

Bernard, P., J. Lecomte, J. Dendien et J.-M. Pierrel. 2002, « Computerized linguistic resources of the research laboratory ATILF for lexical and textual analysis: Frantext, TLFi, and the software Stella. Las Palmas, Espagne (27 mai - 2 juin 2002) », dans *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, p. 1090–1098.

Bernhard, D. 2010, « Apprentissage non supervisé de familles morphologiques : Comparaison de méthodes et aspects multilingues », *Traitement Automatique des Langues*, vol. 51, nº 2, p. 11–39.

Bernhard, D., B. Cartoni et D. Tribout. 2011, « A Task-based Evaluation of French Morphological Resources and Tools », *Linguistic Issues in Language Technology*, vol. 5, nº 2.

Bernhard, D. et I. Gurevych. 2009, « Combining lexical semantic resources with question & answer archives for translation-based answer finding », dans *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL '09)*, Suntec, Singapour, p. 728–736.

Berrendonner, A. 2002, « Les deux syntaxes », *Verbum*, vol. 24, nº 1–2, p. 23–36.

Bescherelle, L.-N. et H. Bescherelle. 1842, *Le Véritable Manuel des conjugaisons, ou la Science des conjugaisons mise à la portée de tout le monde*, Dépôt central des publications classiques, Paris, Royaume de France.

Bickel, B. et J. Nichols. 2005, « Inflectional synthesis of the verb », dans *The World Atlas of Language Structures*, édité par D. G. Martin Haspelmath, Matthew S. Dryer et B. Comrie, Oxford University Press, Oxford, Royaume-Uni, p. 94–97.

Bies, A., J. Mott, C. Warner et S. Kulick. 2012, « English web treebank », cahier de recherche, Linguistic Data Consortium, Philadelphie, Pennsylvanie, États-Unis.

BijanKhan, M. 2004, « The Role of the Corpus in Writing a Grammar: An Introduction to a Software », *Iranian Journal of Linguistics*, vol. 19, nº 2.

Bilhaut, F. et A. Widlöcher. 2006, « Linguastream: An integrated environment for computational linguistics experimentation », dans *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (EACL 2006)*, Trente, Italie, p. 95–98.

Blair, C. R. 1960, « A Program for Correcting Spelling Errors », *Information and Control*, vol. 3, nº 1, p. 60–67.

Blancafort San José, H., G. Recourcé, J. Couto, B. Sagot, R. Stern et D. Teyssou. 2010, « Traitement des inconnus : une approche systématique de l'incomplétude lexicale », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Blevins, J. P. 2006, « Word-based morphology », *Journal of Linguistics*, p. 531–573.

Blevins, J. P. 2016, *Word and Paradigm Morphology*, Oxford University Press, Oxford, Royaume-Uni.

Bloch, O. et W. von Wartburg. 1932, *Dictionnaire étymologique de la langue française*, Presses universitaires de France, Paris, France.

Bloomfield, L. 1933a, *Language*, Holt, Rinchart and Winston, New York City, New York, États-Unis.

Bloomfield, L. 1933b, « A set of postulates for the science of language », *Language*, vol. 3, nº 2, p. 153–164.

Boguraev, B. et T. Briscoe. 1987, « Large lexicons for natural language processing: Utilising the grammar coding system of ldoce », *Computational Linguistics*, vol. 13, nº 3/4, p. 203–218.

Boguraev, B., T. Briscoe, J. Carroll, D. Carter et C. Grover. 1987, « The derivation of a grammatically indexed lexicon from the longman dictionary of contemporary english », dans *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics (ACL '87)*, Stanford, Californie, États-Unis, p. 193–200.

Bohnet, B. 2010, « Very high accuracy and fast dependency parsing is not a contradiction », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède, p. 89–97.

Bonami, O. 1999, *Les constructions du verbe : le cas des groupes prépositionnels argumentaux*, Thèse de doctorat, Université Paris 7.

Bonami, O. 2014, *La structure fine des paradigmes de flexion*, Habilitation à diriger des recherches, Université Denis–Diderot Paris 7.

Bonami, O. et S. Beniamine. 2015, « Implicative structure and joint predictiveness », dans *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*, édité par V. Pirelli, C. Marzi et M. Ferro, Pise, Italie.

Bonami, O. et G. Boyé. 2003, « Supplétion et classes flexionnelles dans la conjugaison du français », *Langages*, vol. 152, p. 102–126.

Bonami, O. et G. Boyé. 2010, « Opaque paradigms, transparent forms in nepali conjugation », Communication orale au Workshop On Theoretical Morphology 5.

Bonami, O., G. Boyé, H. Giraudo et M. Voga. 2008, « Quels verbes sont réguliers en français ? », dans *Actes du 1er Congrès Mondial de Linguistique Française*, Paris, France, p. 1511–1523.

Bonami, O., G. Boyé et F. Henri. 2011, « Measuring inflectional complexity: French and Mauritian », Paper presented at the Workshop on Quantitative Measures in Morphology and Morphological Development.

Bonami, O., G. Caron et C. Plancq. 2014, « Construction d'un lexique flexionnel phonétisé libre du français », dans *Actes du quatrième Congrès Mondial de Linguistique Française*, p. 2583–2596.

Bonami, O. et A. R. Luís. 2013, « A morphologists perspective on Creole complexity », dans *Actes du 19ème Congrès International des Linguistes*, Genève, Suisse.

Bond, F. et K. Paik. 2012, « A survey of wordnets and their licenses », dans *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japon, p. 64–71.

Booij, G. et A. Hulk. 1988, « Introduction », dans *Lexique 7 / Lexique et syntaxe en grammaire générative*, édité par G. Booij et A. Hulk, Presses Universitaires du Septentrion.

Boons, J.-P., A. Guillet et C. Leclère. 1976a, « La structure des phrases simples en français — Classes de constructions transitives », cahier de recherche, LADL, CNRS & Université Paris 7, Paris, France.

Boons, J.-P., A. Guillet et C. Leclère. 1976b, *La structure des phrases simples en français — Constructions intransitives*, Droz, Genève, Suisse.

Booth, B. 1985, « Revising CLAWS », *ICAME news*, vol. 9, p. 29–35.

Borer, H. 2005, *The Normal Course of Events*, Oxford University Press.

Borin, L., M. Forsberg et L. Lönngren. 2008, « The hunting of the BLARK - SALDO, a freely available lexical database for swedish language technology », dans *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, Uppsala University, Uppsala, Suède, p. 21–32.

Bouchard-Côté, A., D. Hall, T. L. Griffiths et D. Klein. 2013, « Automated reconstruction of ancient languages using probabilistic models of sound change », *Proceedings of the National Academy of Sciences*, vol. 110, n° 11, p. 4224–4229.

Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard et Q. Atkinson. 2012, « Mapping the origins and expansion of the Indo-European language family », *Science*, vol. 337, p. 957–960.

Boulanger, J. 2003, *Les inventeurs de dictionnaires : de l'eduba des scribes mésopotamiens au scriptorium des moines médiévaux*, Collection « Regards sur la traduction », Presses de l'Université d'Ottawa, Ottawa, Ontario, Canada.

Boullier, P. 1999, « On Multicomponent TAG parsing », dans *Actes de la 6ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, Cargèse, France, p. 321–326. Cf. aussi *Research Report 3668* at `http://www.inria.fr/RRRT/RR-3668.html`, INRIA-Rocquencourt, France, Apr. 1999, 39 pages.

Boullier, P. 2000, « A cubic time extension of context-free grammars », *Grammars*, vol. 3, n° 2/3, p. 111–131.

Boullier, P., L. Clément, B. Sagot et É. Villemonte de La Clergerie. 2005a, « « simple comme easy :-) » », dans *Actes du workshop EASy de la conférence TALN 2005*, Dourdan, France, p. 57–60.

Boullier, P. et P. Deschamp. 1988–2007, « Le système Syntax ™ – manuel d'utilisation et de mise en œuvre sous Unix™ », Accessible en ligne à l'adresse `http://syntax.gforge.inria.fr/syntax3.8-manual.pdf`.

Boullier, P., A. Nasr et B. Sagot. 2009, « Constructing parse forests that include exactly the *n*-best PCFG trees », dans *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France.

Boullier, P. et B. Sagot. 2005, « Efficient and robust LFG parsing: SxLfg », dans *Proceedings of IWPT'05*, Vancouver, Colombie Britannique, Canada.

Boullier, P. et B. Sagot. 2009a, « Multi-Component Tree Insertion Grammars », dans *Proceedings of the 14th Conference on Formal Grammars (FG 2009)*, Bordeaux, France.

Boullier, P. et B. Sagot. 2009b, « Parsing directed acyclic graphs with range concatenation grammars », dans *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France.

Boullier, P. et B. Sagot. 2010, « Are Very Large Context-Free Grammars Tractable ? », dans *Trends in Parsing Technology*, *Text, Speech and Language Technology*, vol. 43, édité par J. N. Harry Bunt, Paola Merlo, Springer.

Boullier, P., B. Sagot et L. Clément. 2005b, « Un analyseur LFG efficace pour le français : SxLfg », dans *Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France, p. 403–408.

Bourigault, D. et C. Frérot. 2004, « Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène », dans *Actes de la 11ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*, Fès, Maroc, p. 81–90.

Boyd, A. 2009, « Pronunciation modeling in spelling correction for writers of English as a foreign language », dans *Proceedings of the HLT-NAACL'09 Student Research Workshop and Doctoral Consortium*, Boulder, Colorado, États-Unis, p. 31–36.

Boyé, G. 2000, *Problèmes de morpho-phonologie verbale en français, espagnol et italien*, Thèse de doctorat, Université Paris 7.

Boyé, G. 2011, « Régularité et classes flexionnelles dans la conjugaison du français », dans *Des unités morphologiques au lexique*, édité par M. Roché, G. Boyé, N. Hathout, S. Lignon et M. Plénat, Hermès Science, Paris, France.

Brants, S., S. Dipper, S. Hansen, W. Lezius et G. Smith. 2002, « The TIGER treebank », dans *Proceedings of the Workshop on Treebanks and Linguistic Theories*, p. 24–41.

Brants, T. 1996, « Estimating markov model structures », dans *Proceedings of the Fourth Conference on Spoken Language Processing (ICSLP-96*, p. 893–896.

Brants, T. 2000, « TnT: A Statistical Part-of-speech Tagger », dans *Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLC 2000)*, Seattle, Washington, États-Unis, p. 224–231.

Brent, M. R. 1991, « Automatic acquisition of subcategorization frames from untagged text », dans *Proceedings of ACL'91*, p. 209–214.

Bresnan, J. 1982, *The mental representation of grammatical relations*, MIT press, Cambridge, Massachusetts, États-Unis.

Bresnan, J. 2001, *Lexical-Functional Syntax*, Blackwell, Oxford, Royaume-Uni.

Bresnan, J. et S. A. Mchombo. 1995, « The lexical integrity principle: evidence from bantu », *Natural Language and Language Theory*, vol. 13, p. 181–254.

Brill, E. 1995, « Unsupervised learning of disambiguation rules for part of speech tagging », dans *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Massachusetts, États-Unis, p. 1–13.

Brill, E. et R. C. Moore. 2000, « An Improved Error Model for Noisy Channel Spelling Correction », dans *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, Chine.

Briscoe, T. 2001, « From dictionary to corpus to self-organizing dictionary: Learning valency associations in the face of variation and change », dans *Proceedings of the Corpus Linguistics Conference*, Lancaster, Royaume-Uni.

Briscoe, T. et J. Carroll. 1997, « Automatic extraction of subcategorization from corpora », dans *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, D.C., États-Unis.

Brown, D. et R. Evans. 2012, « Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data », dans *Current Issues in Morphological Theory : (Ir)regularity, analogy and frequency*, édité par F. Kiefer, M. Ladányi et P. Siptár, John Benjamins, Amsterdam, Pays-Bas, p. 135–162.

Brown, D. et A. Hippisley. 2012, *Network Morphology: A Defaults-based Theory of Word Structure*, Cambridge University Press, Cambridge, Royaume-Uni.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer et P. S. Roossin. 1990, « A statistical approach to machine translation », *Computational Linguistics*, vol. 16, nº 2, p. 79–85.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra et R. L. Mercer. 1993, « The mathematics of statistical machine translation : Parameter estimation », *Computational Linguistics*, vol. 19, nº 2, p. 263–311.

Buchholz, S. et E. Marsi. 2006, « Conll-x shared task on multilingual dependency parsing », dans *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL X)*, New York City, New York, États-Unis, p. 149–164.

Buck, C. D. 1914, « Is the Suffix of ΒΑΣΙΛΙΣΣΑ, etc., of Macedonian Origin? », *Classical Philology*, vol. 9, nº 4, p. 370–373.

Burnage, G. 1990, « Celex: A Guide for Users », cahier de recherche, University of Nijmegen, Center for Lexical Information.

Butt, M. et T. H. King. 2003, « Grammar writing, testing, and evaluation », dans *Handbook for Language Engineers*, édité par A. Farghaly, CSLI Publications, Stanford, Californie, États-Unis.

de Calmès, M. et G. Pérennou. 1998, « BDLEX: a Lexicon for Spoken and Written French », dans *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, Grenade, Espagne.

Camilleri, M. et G. Walther. 2012, « What small vowels and a large lexicon tell us about maltese verbal inflection », dans *Communication orale au 8ème Colloque des Décembrettes (Décembrettes 8)*, Bordeaux, France.

Can, B. et S. Manandhar. 2009, « Unsupervised learning of morphology by using syntactic categories », dans *Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009)*, Corfou, Grèce.

Candito, M., P. Amsili, L. Barque, F. Benamara, G. De Chalendar, M. Djemaa, P. Haas, R. Huyghe, Y. Y. Mathieu, P. Muller, B. Sagot et L. Vieu. 2014, « Developing a French FrameNet: Methodology and First results », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Candito, M. et B. Crabbé. 2009, « Improving generative statistical parsing with semi-supervised word clustering », dans *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France, p. 138–141.

Candito, M., B. Crabbé, P. Denis et F. Guérin. 2009a, « Analyse syntaxique du français : des constituants aux dépendances », dans *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

Candito, M., B. Crabbé et D. Seddah. 2009b, « On statistical parsing of French with supervised and semi-supervised strategies », dans *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athènes, Grèce.

Candito, M., E. Henestroza Anguiano et D. Seddah. 2011, « A Word Clustering Approach to Domain Adaptation: Effective Parsing of Biomedical Texts », dans *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*, Dublin, Irlande.

Candito, M., J. Nivre, P. Denis et E. H. Anguiano. 2010, « Benchmarking of statistical dependency parsers for french », dans *Proceedings of the 23rd International Conference on Computational Linguistics (CoLing 2010)*, Pékin, Chine, p. 108–116.

Candito, M. et D. Seddah. 2010, « Parsing word clusters », dans *Proceedings of the NAACL/HLT 2010 Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, Californie, États-Unis, p. 76–84.

Candito, M.-H. 1999, *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*, Thèse de doctorat, Université Paris 7.

Carlson, A. et I. Fette. 2007, « Memory-based context-sensitive spelling correction at web scale », dans *Proceedings of ICMLA'07*, Cincinnati, Ohio, États-Unis, p. 166–171.

Carpuat, M. et D. Wu. 2007, « Improving statistical machine translation using word sense disambiguation », dans *The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, p. 61–72.

Cartoni, B. 2009, « Lexical morphology in machine translation: a feasibility study », dans *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*, Athènes, Grèce, p. 130–138.

Casses, B. 2010, « Aligning Wiktionary with Natural Language Processing Resources », .

Chanier, T., C. Poudat, B. Sagot, G. Antoniadis, C. R. Wigham, L. Hriba, J. Longhi et D. Seddah. 2014, « The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres », *JLCL - Journal for Language Technology and Computational Linguistics*, vol. 29, nº 2, p. 1–30.

Charniak, E. 1997, « Statistical parsing with a context-free grammar and word statistics », dans *Proceedings of the 14th National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence (AAAI 1997/IAAI 1997)*, Providence, Rhode Island, États-Unis, p. 598–603.

Charniak, E. et M. Johnson. 2005, « Coarse-to-fine n-best parsing and maxent discriminative reranking », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, États-Unis.

Chiang, D. 2000, « Statistical Parsing with an Automatically-extracted Tree Adjoining Grammar », dans *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, Chine, p. 456–463.

Chomsky, N. 1957, *Syntactic Structures*, Mouton, La Haye, Pays-Bas.

Chomsky, N. 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Massachusetts, États-Unis.

Chomsky, N. 1970, « Remarks on nominalization », dans *Readings in English Transformational Grammar*, édité par R. Jacobs et P. Rosenbaum, Ginn, Waltham, Massachussets, États Unis, p. 184–221.

Chomsky, N. et M. Halle. 1968, *The Sound Pattern of English*, Harper & Row, New York City, New York, États-Unis.

Chrupała, G. 2008, *Towards a machine-learning architecture for lexical functional grammar parsing*, Thèse de doctorat, Dublin City University.

Chrupała, G. 2013, « Text segmentation with character-level text embeddings », dans *Procedings of the ICML Workshop on Deep Learning for Audio, Speech and Lang. Processing*, Atlanta, Géorgie, États-Unis.

Chrupała, G., G. Dinu et J. van Genabith. 2008, « Learning morphology with morfette », dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Chu, Y. J. et T. H. Liu. 1965, « On the Shortest Arborescence of a Directed Graph », *Science Sinica*, vol. 14, p. 1396–1400.

Church, K. 2011, « A pendulum swung too far », *Linguistic Issues in Language Technology*, vol. 6, nº 5.

Church, K. W. 1988, « A stochastic parts program and noun phrase parser for unrestricted text », dans *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLC 1988)*, Austin, Texas, États-Unis, p. 136–143.

Čingienė, J., D. Tcherniak et B. Sagot. 2015, « Sentiment analysis of write-in comments related to organisational change », dans *Proceedings of the 17th Congress of the European Association of Work and Organizational Psychology*, Oslo, Norvège.

Civit, M. et M. A. Martì. 2004, « Building cast3lb: A spanish treebank », *Research on Language and Computation*, vol. 2, nº 4, p. 549 – 574.

Clément, L., B. Sagot et B. Lang. 2004, « Morphology based automatic acquisition of large-coverage lexica », dans *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, p. 1841–1844.

Cohen, P., B. Heeringa et N. Adams. 2002, « An unsupervised algorithm for segmenting categorical timeseries into episodes », *Pattern Detection and Discovery*, p. 117–133.

Collins, M. 1997, « Three Generative, Lexicalised Models for Statistical Parsing », dans *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL 1997)*, Madrid, Espagne, p. 16–23.

Collins, M. 2000, « Discriminative Reranking for Natural Language Parsing », dans *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, San Francisco, Californie, États-Unis.

Collins, M. 2003, « Head-driven statistical models for natural language parsing », *Computational Linguistics*, vol. 29, nº 4, p. 589–637.

Collins, M. et T. Koo. 2005, « Discriminative reranking for natural language parsing », *Computational Linguistics*, vol. 31, nº 1, p. 25–69.

Collobert, R. et J. Weston. 2008, « A unified architecture for natural language processing : Deep neural networks with multitask learning », dans *Proceedings of the 25th International Conference on Machine Learning (ICML 2008)*, Helsinki, Finlande, p. 160–167.

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu et P. Kuksa. 2011, « Natural language processing (almost) from scratch », *Journal of Machine Learning Research*, vol. 12, p. 2493–2537.

Constant, M., M. Candito et D. Seddah. 2013, « The LIGM-Alpage architecture for the SPMRL 2013 Shared Task: Multiword Expression analysis and dependency parsing », dans *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages*, Seattle, Washington, États-Unis, p. 46–52.

Constant, M. et A. Sigogne. 2011, « MWU-aware part-of-speech tagging with a CRF model and lexical resources », dans *Proceedings of the Workshop on Multiword Expressions : From Parsing and Generation to the Real World (MWE 2011)*, Portland, Oregon, États-Unis, p. 49–56.

Constant, M., A. Sigogne et P. Watrin. 2012, « Discriminative strategies to integrate multiword expression recognition and parsing », dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, République de Corée, p. 204–212.

Constant, M. et I. Tellier. 2012, « Evaluating the impact of external lexical resources into a crf-based multiword segmenter and part-of-speech tagger », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, p. 646–650.

Constant, M., I. Tellier, D. Duchier, Y. Dupont, A. Sigogne et S. Billot. 2011, « Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France, p. 321–332.

Copestake, A. 2002, *Implementing Typed Feature Structure Grammars*, CSLI Publications, Stanford, Californie, États-Unis.

Copestake, A. et D. Flickinger. 2000, « An Open Source Grammar Development Environment and Broad-coverage English Grammar Using HPSG », dans *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athènes, Grèce.

Copestake, A., A. Sanfilippo, T. Briscoe et V. de Paiva. 1993, « The ACQUILEX LKB: An Introduction », dans *Inheritance, Defaults and the Lexicon*, édité par T. Briscoe, A. Copestake et V. de Paiva, Cambridge University Press, Cambridge, Royaume-Uni, p. 148–163.

Corbett, G. G. 2003, « Agreement: the range of the phenomenon and the principles of the surrey database of agreement », *Transactions of the philological society*, vol. 101, p. 155–202.

Corbett, G. G. 2006, *Agreement*, Cambridge Textbooks in Linguistics, Cambridge University Press.

Corbett, G. G. 2007, « Canonical typology, suppletion and possible words », *Language*, vol. 83, p. 8–42.

Corbett, G. G. 2009, « Canonical inflectional classes », dans *Selected Proceedings of the 6th Décembrettes : Morphology in Bordeaux*.

Corbett, G. G. 2010, « Morphomic splits », Communication orale à l'atelier "Perspectives on the Morphome" à l'Université de Coimbra.

Corbett, G. G. et N. M. Fraser. 1993, « Network morphology: a DATR account of Russian nominal inflection », *Journal of Linguistics*, vol. 29, p. 113–142.

Coseriu, E. 1964, « Pour une sémantique diachronique structurale », *Travaux de linguistique et de littérature*, vol. 2, n° 1, p. 139–186.

Courtois, B. 1990, « Un système de dictionnaires électroniques pour les mots simples du français », *Langue française*, vol. 87, n° 1, p. 11–22.

Cowan, B. et M. Collins. 2005, « Morphology and reranking for the statistical parsing of spanish », dans *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 795–802.

Crabbé, B. 2014, « An LR-inspired generalized lexicalized phrase structure parser », dans *Proceedings of the 25th International Conference on Computational Linguistics (CoLing 2014)*, Dublin, Irlande.

Crabbé, B. et M. Candito. 2008, « Expériences d'analyses syntaxique statistique du français », dans *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*, Avignon, France.

Crabbé, B., D. Duchier, C. Gardent, J. Le Roux et Y. Parmentier. 2013, « Xmg: eXtensible MetaGrammar », *Computational Linguistics*, vol. 39, nº 3, p. 591–629.

Creutz, M. et K. Lagus. 2005, « Inducing the morphological lexicon of a natural language from unannotated text », dans *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, Finlande, p. 106–113.

Cuadros, M. et G. Rigau. 2006, « Quality assessment of large scale knowledge resources », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Sydney, Australie, p. 534–541.

Cunningham, H., D. Maynard, K. Bontcheva et V. Tablan. 2002, « GATE: A framework and graphical development environment for robust NLP tools and applications », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphie, Pennsylvanie, États-Unis.

Dal, G., F. Namer, N. Hathout et P. Amsili. 1999, « Construire un lexique dérivationnel : théorie et réalisations », dans *Actes de la 6ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 1999)*, Cargèse, France, p. 115–124.

Damerau, F. 1964, « A technique for computer detection and correction of spelling errors », *Communications of the ACM*, vol. 7, nº 3, p. 171–176.

Danlos, L. 2005, « Ilimp: Outil pour repérer les occurrences du pronom impersonnel il », dans *Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France.

Danlos, L., Q. Pradet, L. Barque, T. Nakamura et M. Constant. 2016, « Un Verbənet du français », *Traitement Automatique des Langues*, vol. 57, nº 1, p. 33–58.

Danlos, L. et B. Sagot. 2008, « Constructions pronominales dans dicovalence et le lexique-grammaire — intégration dans le Le*fff* », dans *Proceedings of the 27th Conference on Lexis and Grammar*, L'Aquila, Italie.

Danlos, L. et B. Sagot. 2010, « Ponctuations fortes abusives », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Danlos, L., B. Sagot et S. Salmon-Alt. 2006, « French frozen verbal expressions: from lexicon-grammar tables to NLP applications », dans *Proceedings of the 25th Lexis and Grammar Conference*, Palerme, Italie.

Danlos, L., B. Sagot et R. Stern. 2010, « Analyse discursive des incises de citation », dans *Actes du 2ème Congrès Mondial de Linguistique Française (CMLF 2010)*, La Nouvelle-Orléans, Louisiane, États-Unis.

Das, D. et S. Petrov. 2011, « Unsupervised part-of-speech tagging with bilingual graph-based projections », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, Portland, Oregon, États-Unis, p. 600–609.

Daumé III, H. 2004, « Notes on CG and LM-BFGS optimization of logistic regression », Article disponible à l'adresse `http://pub.hal3.name#daume04cg-bfgs`, implémentation disponible à l'adresse `http://hal3.name/megam/`.

De La Briandais, R. 1959, « File searching using variable length keys », dans *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference (IRE-AIEE-ACM 1959 [Western])*, San Francisco, Californie, États-Unis, p. 295–298.

Declerck, T., A. G. Pérez, O. Vela, Z. Gantner et D. Manzano-Macho. 2006, « Multilingual lexical semantic resources for ontology translation », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

DeFrancis, J. 1984, *The Chinese language: Fact and fantasy*, University of Hawaii Press, Honolulu, Hawaii, États-Unis.

Dellert, J. 2016, « Using Causal Inference to Detect Directional Tendencies in Semantic Evolution », dans *Proceedings of the 11th International Conference on the Evolution of Language (EvoLang 2016)*, La Nouvelle Orléans, Lousiane, États-Unis.

Denis, P. et B. Sagot. 2009, « Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art POS tagging with less human effort », dans *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009)*, Hong Kong, Chine.

Denis, P. et B. Sagot. 2010, « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morphosyntaxique état-de-l'art du français », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Denis, P. et B. Sagot. 2012, « Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging », *Language Resources and Evaluation*, vol. 46, n° 4, p. 721–736.

Diab, M. 2004, « Feasibility of bootstrapping an Arabic WordNet leveraging parallel corpora and an English WordNet », dans *Proceedings of the Arabic Language Resources and Tools conference (NEMLAR)*, Le Caire, Égypte.

Dixon, R. M. W. 1977, « Some Phonological Rules in Yidiɲ », *Linguistic Inquiry*, vol. 8, p. 1–34.

Dixon, R. M. W. et A. Y. Aikhenvald, éd.. 2003, *Word: A Cross-linguistic Typology*, Cambridge University Press, Cambridge, Royaume-Uni.

Dressler, W. 2004, « Degrees of grammatical productivity in inflectional morphology », *Italian Journal of Linguistics*, vol. 15, p. 31–62.

Dressler, W. U. et A. M. Thornton. 1996, « Italian nominal inflection », *Wiener Linguistische Gazette*, vol. 55–57, p. 1–26.

Duanmu, S. 1998, « Wordhood in Chinese », dans *New Approaches to Chinese Word Formation : Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, édité par J. Packard, Mouton de Gruyter, Berlin, Allemagne, p. 135–196.

Dubois, J. et F. Dubois-Charlier. 1997, *Les verbes français*, Larousse-Bordas, Paris, France.

Duong, L., P. Cook, S. Bird et P. Pecina. 2013, « Simpler unsupervised pos tagging with bilingual projections », dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgarie, p. 634–639.

Dyvik, H. 2002, « Translations as semantic mirrors: from parallel corpus to wordnet », dans *Post-proceedings of the ICAME 2002 Conference (revised version)*, Göteborg, Suède.

Earley, J. 1968, *An Efficient Context-free Parsing Algorithm*, Thèse de doctorat, Department of Computer Science, Carnegie-Mellon University.

Earley, J. 1970, « An efficient context-free parsing algorithm », *Communications of the ACM*, vol. 13, n° 2, p. 94–102.

Edmonds, J. 1967, « Optimum Branchings », *Journal of Research of the National Bureau of Standards*, vol. 71B, p. 233–240.

Ehret, K. 2014, « Kolmogorov Complexity of Morphs and Constructions in English », *LiLT*, vol. 11, n° 2, p. 43–71.

Elsner, M. et E. Charniak. 2011, « Disentangling chat with local coherence models », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, Portland, Oregon, États-Unis, p. 1179–1189.

Emerson, T. 2005, « The second international chinese word segmentation bakeoff », dans *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, vol. 133.

Engel, U. 2009, *Deutsche Grammatik, 4., völlig neu bearbeitete Auflage*, *Grundlagen der Germanistik*, vol. 22, Erich Schmidt Verlag, Berlin, Allemagne.

Erjavec, T. 2010, « Multext-east version 4: Multilingual morphosyntactic specifications, lexicons and corpora », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Erjavec, T. et D. Fišer. 2006, « Building Slovene WordNet », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Evans, N., J. Fletcher et B. Ross. 2008, « Big words, small phrases: Mismatches between pause units and the polysynthetic word in Dalabon », *Linguistics*, vol. 46, n° 1, p. 89–129.

Evans, R. P. et G. Gazdar. 1989, « Inference in datr », dans *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, p. 66–71.

van den Eynde, K. et P. Mertens. 2003, « La valence : l'approche pronominale et son application au lexique verbal », *French Language Studies*, vol. 13, nº 1, p. 63–104.

van den Eynde, K. et P. Mertens. 2006, « Valency dictionary - DICOVALENCY: user's guide », `http://bach.arts.kuleuven.be/dicovalence/`.

Falk, I., D. Bernhard, C. Gérard et R. Potier-Ferry. 2014, « Étiquetage morpho-syntaxique pour des mots nouveaux », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France.

Feldman, A., J. Hana et C. Brew. 2006, « A cross-language approach to rapid creation of new morpho-syntactically annotated resources », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, p. 549–554.

Fellbaum, C., éd.. 1998, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, Massachusetts, États-Unis.

Fernandez, M., É. Villemonte de La Clergerie et M. Vilares. 2007, « Knowledge acquisition through error-mining », dans *Proceedings of the 6th conference on Recent Advances in Natural Language Processing (RANLP 2007)*, Borovets, Bulgarie.

Fillmore, C. 1968, « The case for case », dans *Universals in Linguistic Theory*, édité par E. Bach et R. T. Harms, Holt, Rinehart and Winston, New York City, New York, États-Unis, p. 1–88.

Fillmore, C. J. 1982, « Frame semantics », dans *Linguistics in the Morning Calm*, Hanshin Publishing Co., Séoul, République de Corée, p. 111–137.

Fillmore, C. J. 2006, « Frame semantics and the nature of language », *Annals of the New York Academy of Sciences*, vol. 280, nº 1, p. 20–32.

Finkel, R. et G. T. Stump. 2002, « Generating hebrew verb morphology by default inheritance hierarchies », dans *Proceedings of the ACL 2002 Workshop on Computational Approaches to Semitic Languages*, Philadelphie, Pennsylvanie, États-Unis.

Fišer, D. 2007, « Leveraging parallel corpora and existing wordnets for automatic construction of the slovene wordnet », dans *Proceedings of the 3rd Language and Technology Conference (LTC 2007)*, Poznań, Pologne.

Fišer, D. et N. Ljubešić. 2011, « Bilingual lexicon extraction from comparable corpora for closely related languages. », dans *Proceedings of the 8th conference on Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgarie, p. 125–131.

Fišer, D. et B. Sagot. 2008, « Combining Multiple Resources to Build Reliable Wordnets », dans *Proceedings of the 11th International Conference on Text, Speech and Dialogue (TSD 2008)*, Brno, République tchèque.

Fišer, D. et B. Sagot. 2015, « Constructing a poor man's wordnet in a resource-rich world », *Language Resources and Evaluation*, p. 1–35.

Flobert, P. 1967, « Déponent et passif en italique et en celtique », *Annales de Bretagne*, vol. 74, nº 4, p. 567–604.

Forsberg, M., H. Hammarström et A. Ranta. 2006, « Morphological lexicon extraction from raw text data », dans *Proceedings of FinTAL 2006, LNAI 4139*, Springer-Verlag, Turku, Finlande, p. 488–499.

Fort, K., G. Adda, B. Sagot, J. Mariani et A. Couillault. 2014, « Crowdsourcing for Language Resource Development : Criticisms About Amazon Mechanical Turk Overpowering Use », dans *Human Language Technology Challenges for Computer Science and Linguistics*, *Lecture Notes in Computer Science*, vol. 8387, édité par Z. Vetulani et J. Mariani, Springer International Publishing, p. 303–314.

Fort, K. et B. Sagot. 2010, « Influence of pre-annotation on POS-tagged corpus development », dans *Proceedings of the Fourth ACL Linguistic Annotation Workshop (LAW IV)*, Uppsala, Suède, p. 56–63.

Fortescue, M. 1999, « The rise and fall of polysynthesis in the eskimo-aleut family », *Sprachtypologie und Universalienforschung*, vol. 52, nº 3/4, p. 282–297.

Foster, J. 2010, « "cba to check the spelling" : Investigating parser performance on discussion forum posts », dans *Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT 2010)*, Los Angeles, Californie, États-Unis, p. 381–384.

Foster, J., Ö. Çetinoğlu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan et J. van Genabith. 2011a, « #hardtoparse: Pos tagging and parsing the twitterverse », dans *Proceedings of the AAAI 2011 Workshop On Analyzing Microtext*, San Francisco, Californie, États-Unis.

Foster, J., Ö. Çetinoğlu, J. Wagner, J. Le Roux, J. Nivre, D. Hogan et J. van Genabith. 2011b, « From news to comment: Resources and benchmarks for parsing the language of web 2.0 », dans *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thaïlande, p. 893–901.

Francis, W. N. et H. Kucera. 1964, *Manual of Information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*, Brown University, Providence, Rhode Island, États-Unis.

Francopoulo, G., M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet et C. Soria. 2006, « Lexical Markup Framework (LMF) », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Fredkin, E. 1960, « Trie memory », *Communications of the ACM*, vol. 3, nº 9, p. 490–499.

Fung, P. 1995, « A pattern matching method for finding noun and proper noun translations from noisy parallel corpora », dans *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL 1995)*, Cambridge, Massachusetts, États-Unis, p. 236–243.

Fung, P. 1998, « A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora », dans *Machine Translation and the Information Soup, Revised Selected Papers from the Third Conference of the Association for Machine Translation in the Americas (AMTA '98)*, *Lecture Notes in Computer Science (LNCS)*, vol. 1529, Springer-Verlag, Langhorne, Pennsylvanie, États-Unis, p. 1–17.

Gábor, K., M. Apidianaki, B. Sagot et É. Villemonte de La Clergerie. 2012, « Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie.

Gábor, K. et B. Sagot. 2014, « Automated Error Detection in Digitized Cultural Heritage Documents », dans *Proceedings of the EACL 2014 Workshop on Language Technology for Cultural Heritage*, Göteborg, Suède.

Gabrilovich, E. et S. Markovitch. 2006, « Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge », dans *Proceedings of the 21st national conference on Artificial intelligence (AAAI'06)*, AAAI Press, Boston, Massachusetts, États-Unis, p. 1301–1306.

Gaiffe, B. et K. Nehbi. 2009, « Le corpus de l'Est Républicain », cahier de recherche, Atilf. Http ://www.cnrtl.fr/corpus/estrepublicain/.

Gala, N., V. Rey et M. Zock. 2010, « A tool for linking stems and conceptual fragments to enhance word access », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

García-Miguel, J. M. et F. J. Albertuz. 2005, « Verbs, semantic classes and semantic roles in the ADESSE project », dans *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*.

Garnier, R. 2016, *La dérivation inverse en latin*, 157, Innsbrucker Beiträge zur Sprachwissenschaft.

Garnier, R., L. Sagart et B. Sagot. 2017, « Milk and the Indo-Europeans », dans *Language Dispersal Beyond Farming*, édité par M. Robeets et A. Savalyev, John Benjamins Publishing Company, p. 291–311.

Garnier, R. et B. Sagot. 2015, « Could Greek and Italic share a same Indo-European substratum ? », dans *Proceedings of the 22nd International Conference on Historical Linguistics (ICHL 2015)*, Naples, Italie.

Garnier, R. et B. Sagot. 2017, « A shared substrate between Greek and Italic », *Indogermanische Forschungen*, vol. 122, n° 1, p. 29–60.

Garnier, R. et B. Sagot. 2018a, « Metathesis of Proto-Indo-European Sonorants », *Münchener Studien zur Sprachwissenschaft*. à paraître.

Garnier, R. et B. Sagot. 2018b, « New results on a centum substratum in Greek : the Lydian connection », dans *International Colloquium on Loanwords and Substrata in Indo-European languages*, Limoges, France.

Garrett, A. J. 2008, « Paradigmatic uniformity and markedness : Historical convergence and universal grammar », dans *Explaining linguistic universals*, édité par J. Good, Oxford University Press, p. 125–143.

Gaussier, É. 1999, « Unsupervised learning of derivational morphology from inflectional lexicons », dans *Proceedings of the workshop on Unsupervised Learning in Natural Language Processing*, College Park, Maryland, États-Unis.

Gazdar, G. 1985, *Generalized Phrase Structure Grammar*, Harvard University Press, Cambridge, Massachusetts, États-Unis.

Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan et N. A. Smith. 2011, « Part-of-speech tagging for twitter: Annotation, features, and experiments », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, Portland, Oregon, États-Unis, p. 42–47.

Goldberg, A. E. 1995, *Constructions: A Construction Grammar Approach to Argument Structure*, Cognitive Theory of Language and Culture Series, University of Chicago Press, Chicago, Illinois, États-Unis.

Goldberg, Y. 2017, *Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies)*, Morgan & Claypool.

Goldberg, Y. et M. Elhadad. 2013, « Word segmentation, unknown-word resolution, and morphological agreement in a hebrew parsing system », *Computational Linguistics*, vol. 39, nº 1, p. 121–160.

Goldberg, Y. et R. Tsarfaty. 2008, « A single generative model for joint morphological segmentation and syntactic parsing », dans *Proceedings of ACL*, p. 371–379.

Goldberg, Y., R. Tsarfaty, M. Adler et M. Elhadad. 2009, « Enhancing unlexicalized parsing performance using a wide coverage lexicon, fuzzy tag-set mapping, and em-hmm-based lexical probabilities », dans *Proceedings of the 12th Conference of the European Chapter of the ACL*, p. 327–335.

Goldsmith, J. 2001, « Unsupervised learning of the morphology of a natural language », *Computational Linguistics*, vol. 27, nº 2, p. 153–198.

Goldwater, S. et T. L. Griffiths. 2007, « A fully Bayesian approach to unsupervised part-of-speech tagging », dans *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, République tchèque, p. 744–751.

Grad, A. et H. Leeming. 1999, *Slovene–English Dictionary*, Državna založba Slovenije, Ljubljana, Slovénie.

Grad, A., R. Škerlj et N. Vitorovič. 1999, *English–Slovene Dictionary*, Državna založba Slovenije, Ljubljana, Slovénie.

Gray, R. D. et Q. D. Atkinson. 2003, « Language-tree divergence times support the Anatolian theory of Indo-European origin », *Nature*, vol. 426, p. 435–439.

Green, S., M.-C. de Marneffe et C. D. Manning. 2013, « Parsing models for identifying multiword expressions », *Computational Linguistics*, vol. 39, nº 1, p. 195–227.

Greene, B. et G. M. Rubin. 1971, « Automatic grammatical tagging of English », cahier de recherche, Department of Linguistics, Brown University, Providence, Rhode Island, États-Unis.

Grishman, R., C. Macleod et A. Meyers. 1994, « Comlex syntax: Building a computational lexicon », dans *Proceedings of the 15th International Conference on Computational Linguistics (CoLing 1994)*, Kyoto, Japon, p. 268–272.

Gross, G. 1996, *Les expressions figées en français : noms composés et autres locutions*, Collection L'essentiel français, Ophrys, Paris, France.

Gross, M. 1975, *Méthodes en syntaxe : Régimes des constructions complétives*, Hermann, Paris, France.

Gui, T., Q. Zhang, H. Huang, M. Peng et X. Huang. 2017, « Part-of-Speech Tagging for Twitter with Adversarial Neural Networks », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhague, Danemark, p. 2411–2420.

Guillaume, B. et G. Perrier. 2010a, « Interaction grammars », *Research on Language and Computation*.

Guillaume, B. et G. Perrier. 2010b, « LEOPAR, un analyseur syntaxique pour les grammaires d'interaction », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Guillet, A. et C. Leclère. 1992, *La structure des phrases simples en français — Les constructions transitives locatives*, Droz, Genève, Suisse.

Guimier de Neef, E., A. Debeurme et J. Park. 2007, « TiLT correcteur de SMS : évaluation et bilan qualitatif », dans *Actes de la 14ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, Toulouse, France, p. 123–132.

Hajič, J. 2000, « Morphological Tagging: Data vs. Dictionaries », dans *Proceedings of ANLP'00*, Seattle, Washington, États-Unis, p. 94–101.

Hajič, J., A. Böhmová, E. Hajičová et B. Vidová Hladká. 2000, « The Prague Dependency Treebank: A Three-level Annotation Scenario », dans *Treebanks : Building and Using Parsed Corpora*, édité par A. Abeillé, Kluwer Academic Publisher, Dordrecht, Pays-Bas, p. 103–127.

Hajič, J., J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, et M. ševčiková Razímová. 2006, « Prague Dependency Treebank 2.0 », http ://ufal.mff.cuni.cz/pdt2.0/.

Hakkani-Tür, D. Z., K. Oflazer et G. Tür. 2002, « Statistical morphological disambiguation for agglutinative languages », *Computers and the Humanities*, vol. 36, nº 4, p. 381–410.

Hall, D., D. Jurafsky et C. D. Manning. 2008, « Studying the history of ideas using topic models », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Honolulu, Hawaii, États-Unis, p. 363–371.

Hall, T. A. et U. Kleinhenz, éd.. 1999, *Studies on the Phonological Word*, n° 174 dans Current Issues in Linguistic Theories, John Benjamins B.V.

Halle, M. et A. Marantz. 1993, « Distributed morphology and the pieces of inflection », dans *The view from building 20*, édité par K. Hale et S. J. Keyser, MIT Press, Cambridge, Massachusetts, États-Unis, p. 111–176.

Hamon, O., D. Mostefa, C. Ayache, P. Paroubek, A. Vilnat et É. Villemonte de La Clergerie. 2008, « PASSAGE: from French parser evaluation to large sized treebank », dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Hankamer, J. 1989, « Morphological parsing and the lexicon », dans *Lexical Representation and Process*, édité par W. Marslen-Wilson, MIT Press, Cambridge, Massachusetts, États-Unis, p. 392–408.

Hanoka, V. 2015, *Extraction et complétion de terminologies multilingues*, Thèse de doctorat, Université Denis–Diderot Paris 7.

Hanoka, V. et B. Sagot. 2012, « Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie.

Harabagiu, S., D. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. Bunescu, R. Girju, V. Rus et P. Morarescu. 2000, « Falcon: Boosting knowledge for answer engines », dans *Proceedings of TREC-9*, p. 479–488.

Harris, Z. S. 1955, « From phoneme to morpheme », *Language*, vol. 31, n° 2, p. 190–222.

Harris, Z. S. 1962, *String Analysis of Sentence Structure*, Mouton, La Haye, Pays-Bas.

Haspelmath, M. 2011, « The indeterminacy of word segmentation and the nature of morphology and syntax », *Folia Linguistica*, vol. 45, n° 1, p. 31–80.

Haspelmath, M. et A. Sims. 2010, *Understanding Morphology, 2nd ed.*, Understanding Language, Taylor & Francis, Londres, Royaume-Uni.

Hathout, N. 2010, « Morphonette: a morphological network of French », Accessible à l'adresse `https://arxiv.org/abs/1005.3902`.

Hathout, N., F. Sajous et B. Calderone. 2014, « GLÀFF, a large versatile French lexicon », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Hathout, N. et L. Tanguy. 2005, « WEBAFFIX : une boîte à outils d'acquisition lexicale à partir du Web », *Revue Québécoise de Linguistique*, vol. 32, n° 1, p. 61–84.

Hauer, B. et G. Kondrak. 2011, « Clustering semantically equivalent words into cognate sets in multilingual lists », dans *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thaïlande, p. 865–873.

Henestroza Anguiano, E. et M. Candito. 2011, « Parse correction with specialized models for difficult attachment types », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Édimbourg, Royaume-Uni.

Hewlett, D. et P. Cohen. 2011, « Fully unsupervised word segmentation with BVE and MDL », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers*, vol. 2, p. 540–545.

Hill, N. W. et J.-M. List. 2017, « Challenges of annotation and analysis in computer-assisted language comparison : a case study on Burmish languages », *Yearbook of the Poznań Linguistic Meeting*, vol. 3, nº 1.

Hinrichs, E. et H. Telljohann. 2009, « Constructing a Valence Lexicon for a Treebank of German », dans *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, Groningue, Pays-Bas.

Hochreiter, S. et J. Schmidhuber. 1997, « Long short-term memory », *Neural Computation*, vol. 9, nº 8, p. 1735–1780.

Hockett, C. F. 1954, « Two models of linguistc descriptions », *Words*, vol. 10, p. 210–234.

Hopcroft, J. D. et J. E. Ullman. 1979, *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley, Reading, Massachusetts, États-Unis.

Horsmann, T. et T. Zesch. 2017, « Do LSTMs really work so well for PoS tagging ? – A replication study », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, Copenhague, Danemark.

Huang, C.-R., K.-J. Chen et L.-L. Chang. 1996, « Segmentation Standard for Chinese Natural Language Processing », dans *Proceedings of the 16th International Conference on Computational Linguistics (CoLing 1996)*, Copenhague, Danemark, p. 1045–1048.

Huang, C.-T. J. 1984, « Phrase structure, lexical integrity, and Chinese compounds », *Journal of the Chinese Language Teachers Association*, vol. 19, nº 2, p. 53–78.

Huang, L. et K. Sagae. 2010, « Dynamic programming for linear-time incremental parsing », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède, p. 1077–1086.

Huet, G. 2005, « A functional toolkit for morphological and phonological processing, application to a sanskrit tagger », *Journal of Functional Programming*, vol. 15, nº 4, p. 573–614.

Huet, G. 2009, « Sanskrit Segmentation », dans *Communication orale à la 28ème South Asian Languages Analysis Roundtable*, Denton, Texas, États-Unis.

Hulden, M. 2009, « Foma: a finite-state compiler and library », dans *Proceedings of EACL (demos)*, p. 29–32.

Ide, N., T. Erjavec et D. Tufiş. 2002, « Sense discrimination with parallel corpora », dans *Proceedings of the ACL 2002 workshop on Word sense disambiguation : recent successes and future directions (WSD 2002)*, Philadelphie, Pennsylvanie, États-Unis, p. 61–66.

Ide, N. et J. Véronis. 1994, « MULTEXT: Multilingual text tools and corpora », dans *Proceedings of the 15th International Conference on Computational Linguistics (CoLing 1994)*, Kyoto, Japon.

Inkpen, D., O. Frunza et G. Kondrak. 2005, « Automatic identification of cognates and false friends in French and English », dans *Proceedings of the 5th conference on Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgarie, p. 251–257.

Inoue, G., H. Shindo et Y. Matsumoto. 2017, « Joint prediction of morphosyntactic categories for fine-grained arabic part-of-speech tagging exploiting tag dictionary information », dans *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Vancouver, Colombie Britannique, Canada, p. 421–431.

Jackendoff, R. 1990, *Semantic Structures*, MIT Press, Cambridge, Massachusetts, États-Unis.

Jacques, G., A. Lahaussois, B. Michailovsky et D. B. Rai. 2012, « An overview of Khaling verbal morphology », *Language and Linguistics*, vol. 13.6, p. 1095–1170.

Jacquin, C., E. Desmontils et L. Monceaux. 2007, « French EuroWordNet lexical database improvements », dans *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2007)*, Mexico, Mexique, p. 12–22.

Jäger, G., J.-M. List et P. Sofroniev. 2017, « Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists », dans *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valence, Espagne, p. 1204–1215.

Jespersen, O. 1924, *Philosophy of Grammar*, George Allen & Unwin, Londres, Royaume-Uni.

Jin, Z. et K. Tanaka-Ishii. 2006, « Unsupervised segmentation of Chinese text by use of branching entropy », dans *Proceedings of the COLING/ACL on Main conference poster sessions*, p. 428–435.

Johansson, S., G. N. Leech et H. Goodluck. 1978, *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computer*, Department of English, University of Oslo, Oslo, Norvège.

Johnson, M. 1998, « PCFG Models of Linguistic Tree Representations », *Computational Linguistics*, vol. 24, n° 4, p. 613–632.

Joseph, B. D. 2001, « Defining “Word” in Modern Greek : A Response to Philippaki-Warburton & Spyropoulos 1999 », dans *Yearbook of Morphology*, édité par G. Booij et J. van Marle, p. 87–114.

Joshi, A. K. 1987, « An introduction to Tree Adjoining Grammars », dans *Mathematics of Language*, édité par A. Manaster-Ramer, John Benjamins, Amsterdam, Pays-Bas, p. 87–114.

Joshi, A. K. et P. Hopely. 1996, « A parser from antiquity », *Natural Language Engineering*, vol. 2, n° 4, p. 291–294.

Joshi, A. K., L. S. Levy et M. Takahashi. 1975, « Tree adjunct grammars », *Journal of Computer and System Sciences*, vol. 10, n° 1, p. 136–163.

Juola, P. 1998, « Measuring linguistic complexity: The morphological tier », *Journal of Quantitative Linguistics*, vol. 5, n° 3, p. 206–13.

Juola, P. 2008, « Assessing Linguistic Complexity », dans *Language Complexity : Typology, Contact, Change*, édité par M. Miestamo, K. Sinnemäki et F. Karlsson, John Benjamins Press, Amsterdam, Pays-Bas.

Kageura, K. et B. Umino. 1996, « Methods of automatic term recognition — a review », *Terminology*, vol. 3, n° 2, p. 259–289.

Kahane, S. 2001, « Grammaires de dépendance formelles et théorie Sens-Texte », dans *Actes de la 8ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, Tours, France.

Kahane, S. 2008, « Les unités minimales de la syntaxe et de la sémantique : le cas du français », dans *Actes du 1er Congrès Mondial de Linguistique Française (CMLF 2008)*, Paris, France. Version corrigée 16/07/08, accessible sur la page internet de l'auteur.

Kallmeyer, L. et W. Maier. 2010, « Data-driven parsing with probabilistic linear context-free rewriting systems », dans *Proceedings of the 23rd International Conference on Computational Linguistics (CoLing 2010)*, Pékin, Chine, p. 537–545.

Kallmeyer, L. et W. Maier. 2013, « Data-driven parsing using probabilistic linear context-free rewriting systems », *Computational Linguistics*, vol. 39, n° 1, p. 87–119.

Kaplan, R. et J. Bresnan. 1982, « Lexical-functional grammar: a formal system for grammatical representation », dans *The Mental Representation of Grammatical Relations*, édité par J. Bresnan, MIT Press, Cambridge, Massachusetts, États-Unis, p. 173–281.

Kaplan, R. M. et I. John T. Maxwell. 1993–1996, « LFG Grammar Writer's Workbench », `ftp://ftp.parc.xerox.com/pub/lfg/lfgmanual.ps`.

Kaplan, R. M. et M. Kay. 1994, « Regular models of phonological rule systems », *Computational Linguistics*, vol. 20, n° 3, p. 331–378.

Karanasou, P. et L. Lamel. 2011, « Pronunciation variants generation using smt-inspired approaches », dans *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, République tchèque, p. 4908–4911.

Karttunen, L. 2003, « Computing with realizational morphology », dans *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science (LNCS)*, vol. 2588, édité par A. Gelbukh, Springer-Verlag, Mexico, Mexique, p. 203–214.

Karttunen, L., J.-P. Chanod, G. Grefenstette et A. Schille. 1996, « Regular expressions for language engineering », *Natural Language Engineering*, vol. 2, nº 4, p. 305–328.

Kasami, T. 1965, « An efficient recognition and syntax algorithm for context-free languages », cahier de recherche AFCRL-65-758, Air Force Cambridge Research Laboratory, Cambridge, Massachusetts, États-Unis.

Kathol, A. 1995, *Linearization-based German Syntax*, Thèse de doctorat, Ohio State University.

Kay, M. 1980, « Algorithm schemata and data structures in syntactic processing », cahier de recherche CSL-80-12, Xerox Palo Alto Research Center, Palo Alto, Californie, États-Unis.

Kempe, A. 1999, « Experiments in unsupervised entropy-based corpus segmentation », dans *Workshop of EACL in Computational Natural Language Learning*, p. 7–13.

Kernighan, M. D., K. W. Church et W. A. Gale. 1990, « A Spelling Correction Program Based on a Noisy Channel Model », dans *Proceedings of the 13th Conference on Computational Linguistics (COLING 1990)*, Helsinki, Finlande, p. 205–210.

Kibrik, A. E. 2001, « Archi (Caucasian Daghestanian) », dans *The Handbook of Morphology*, édité par A. Spencer et A. M. Zwicky, Blackwell Handbooks in Linguistics, Oxford, Royaume-Uni.

Kilani-Schoch, M. et W. U. Dressler. 2005, *Morphologie naturelle et flexion du verbe français*, Gunter Narr Verlag, Tübingen, Allemagne.

Kim, J.-D., S.-Z. Lee et H.-C. Rim. 1999, « HMM Specialization with Selective Lexicalization », dans *Proceedings of the join SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 1999)*.

Kingsbury, P. et M. Palmer. 2002, « From Treebank to PropBank », dans *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*.

Kinyon, A. 2000, « Hypertags », dans *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Sarrebruck, Allemagne, p. 446–452.

Kiparsky, P. 1982, « Word-formation and the Lexicon », dans *Proceedings of the Mid-America Linguistics Conference*, édité par F. Ingemann, Lawrence, Kansas, États-Unis.

Kipper, K., H. T. Dang et M. Palmer. 2000, « Class-based construction of a verb lexicon », dans *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, p. 691–696.

Kipper Schuler, K. 2005, *VerbNet: A broad-coverage, comprehensive verb lexicon*, Thèse de doctorat, University of Pennsylvania.

Kirkpatrick, B. 1987, *Roget's Thesaurus of English Words and Phrases*, Penguin reference books, Penguin, Londres, Royaume-Uni.

Klein, D. et C. D. Manning. 2003, « Accurate unlexicalized parsing », dans *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japon, p. 423–430.

Klein, S. et R. F. Simmons. 1963, « A Computational Approach to Grammatical Coding of English Words », *Journal of the ACM*, vol. 10, nº 3, p. 334–347.

Knight, K. et S. K. Luk. 1994, « Building a large-scale knowledge base for machine translation », dans *Proceedings of the twelfth national conference on Artificial intelligence (AAAI '94)*, Seattle, Washington, États-Unis, p. 773–778.

Kobus, C., F. Yvon et G. Damnati. 2008, « Transcrire les SMS comme on reconnaît la parole », dans *Actes de la 15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*, Avignon, France, p. 128–138.

Koehl, A. 2013, « Une base de données des noms désadjectivaux du français : le modèle mordan », dans *Proceedings of Corpus et Outils en Linguistique, langues et parole*, Strasbourg, France.

Koehn, P. et K. Knight. 2002, « Learning a translation lexicon from monolingual corpora », dans *Proceedings of the ACL 2002 Workshop on Unsupervised Lexical Acquisition (SIGLEX 2002)*, Philadelphie, Pennsylvanie, États-Unis, p. 9–16.

Kolmogorov, A. N. 1963, « On tables of random numbers », *Sankhyā : The Indian Journal of Statistics, Series A*, vol. 25, nº 4, p. 369–376.

Kolmogorov, A. N. 1965, « Three approaches to the quantitative defintion of information », *Problems in Information Transmission*, vol. 1, p. 1–7.

Kondrak, G. 2009, « Identification of Cognates and Recurrent Sound Correspondences in Word Lists », *Traitement Automatique des Langues*, vol. 50, nº 2, p. 201–235.

Kondrak, G. et B. Dorr. 2004, « Identification of confusable drug names: A new approach and evaluation methodology », dans *Proceedings of the 20th International Conference on Computational Linguistics (CoLing 2004)*, Genève, Suisse, p. 952–958.

Kong, L., N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer et N. A. Smith. 2014, « A dependency parser for tweets », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, p. 1001–1012.

Koo, T., X. Carreras et M. Collins. 2008, « Simple semi-supervised dependency parsing », dans *Proceedings the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*, Columbus, Ohio, États-Unis, p. 595–603.

Korhonen, A., G. Gorrell et D. McCarthy. 2000, « Statistical filtering and subcategorization frame acquisition », dans *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (EMNLP'00)*, Hong Kong, Chine, p. 199–206.

Korhonen, A., Y. Krymolowski et T. Briscoe. 2006, « A Large Subcategorization Lexicon for Natural Language Processing Applications », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Koskenniemi, K. 1984, « A general computational model for word-form recognition and production », dans *Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics (ACL 1984)*, Stanford, Californie, États-Unis, p. 178–181.

Kočourek, R. 2001, *Essais de linguistique française et anglaise : mots et termes, sens et textes*, n° 48 dans Bibliotheque de l'Information Grammaticale, Peeters.

Kratochvíl, P. 1967, « Modern standard chinese », *Lingua*, vol. 17, n° 1–2, p. 129–152.

Kukich, K. 1992, « Techniques for Automatically Correcting Words in Text », *ACM Computing Surveys*, vol. 24, n° 4, p. 377–439.

Kuno, S. et A. G. Oettinger. 1963, « Multiple-path syntactic analyzer », *Information Processing*, vol. 62, p. 306–312.

Kupść, A. 2007, « Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré », dans *Actes de la 14ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2007)*, Toulouse, France.

Kupść, A. 2008, « Adjectives in TreeLex », dans *Proceedings of the 16th International Conference Intelligent Information Systems*, Zakopane, Pologne, p. 287–296.

Kuryłowicz, J. 1945, « La nature des procès dits « analogiques » », *Acta Linguistica*, vol. 5, n° 1, p. 15–37.

Lafferty, J. D., A. McCallum et F. C. N. Pereira. 2001, « Conditional random fields: Probabilistic models for segmenting and labeling sequence data », dans *ICML*, p. 282–289.

Landauer, T., P. Foltz et D. Laham. 1998, « An introduction to latent semantic analysis », *Discourse processes*, vol. 25, p. 259–284.

Lang, B. 1974, « Deterministic techniques for efficient non-deterministic parsers », dans *Automata, Languages and Programming*, Springer-Verlag, Berlin, République Fédérale d'Allemagne, p. 255–269.

Langacker, R. W. 1972, *Fundamentals of linguistic analysis*, Harcourt Brace Jovanovich, New York City, New York, États-Unis.

Langer, S., P. Maier et J. Oesterle. 1996, « CISLEX, an electronic dictionary for German. Its structure and a lexicographic application », dans *Proceedings of COMPLEX 1996*, Budapest, Hongrie.

Laporte, E., E. Tolone et M. Constant. 2013, « Conversion of Lexicon-Grammar tables to LMF: application to French », dans *LMF : Lexical Markup Framework, theory and practice*, édité par G. Francopoulo, chap. 11, Hermès Science, Paris, France.

Lavallée, J. F. et P. Langlais. 2009, « Unsupervised morphological analysis by formal analogy », dans *Proceedings of the 10th Workshop of the Cross-Language Evaluation Forum (CLEF 2009)*, Corfou, Grèce, p. 618–625.

Lavallée, J.-F. et P. Langlais. 2011, « Moranapho : un système multilingue d'analyse morphologique fondé sur l'analogie formelle », *Traitement Automatique des Langues*, vol. 52, nº 2, p. 17–44.

Lavelli, A. et A. Corazza. 2009, « The berkeley parser at the evalita 2009 constituency parsing task », dans *EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*, Reggio d'Émilie, Italie.

Lavie, A., A. Itai et U. Ornan. 1990, « On the applicability of two level morphology to the inflection of Hebrew verbs », dans *Proceedings of the 15th International Conference on Literary and Linguistic Computing (CLLC 1988)*, Jérusalem, Israël, p. 246–260.

Le Roux, J., M. Constant et A. Rozenknop. 2014, « Syntactic Parsing and Compound Recognition via Dual Decomposition : Application to French », dans *Proceedings of the 25th International Conference on Computational Linguistics (CoLing 2014)*, Dublin, Irlande, p. 1875–1885.

Le Roux, J., J. Foster, J. Wagner, S. Z. Kaljahi, Rasul et A. Bryl. 2012a, « DCU-Paris13 Systems for the SANCL 2012 Shared Task », dans *Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL 2012), a NAACL/HLT 2012 workshop*, Montréal, Québec, Canada, p. 1–4.

Le Roux, J., B. Sagot et D. Seddah. 2012b, « Statistical parsing of spanish and data driven lemmatization », dans *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages (SP-Sem-MRL 2012)*, Jeju, République de Corée, p. 6 pages.

Lee, J. et J. A. Goldsmith. 2013, « Automatic morphological alignment and clustering », Presented at the 2nd American International Morphology Meeting.

Leech, G., R. Garside et E. Atwell. 1983, « The automatic grammatical tagging of the LOB corpus », *ICAME news*, vol. 7, p. 13–33.

Lepage, Y. 1998, « Solving analogies on words: An algorithm », dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, Montréal, Québec, Canada, p. 728–735.

Lepage, Y. 2000, « Languages of analogical strings », dans *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Sarrebruck, Allemagne, p. 488–494.

Levenshtein, V. I. 1965, « Двоичные коды с исправлением выпадений, вставок и замещений символов (Codes binaires permettant la correction de suppressions, d'insertions et d'inversions) », *Доклады Академий Наук СССР (Doklady Akademii Nauk SSSR)*, vol. 163, nº 4, p. 845–848.

Levenshtein, V. I. 1966, « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *Soviet Physics–Doklady*, vol. 10, nº 8, p. 707–710. Traduction de (Levenshtein, 1965).

Levin, B. 1993, *English Verb Classes and Alternations: A Preliminary Investigation*, University of Chicago Press, Chicago, Illinois, États-Unis.

Lieber, R. 1981, *On the Organization of the Lexicon*, Thèse de doctorat, University of New Hamsphire.

Ling, W., T. Luís, L. Marujo, R. F. Astudillo, S. Amir, C. Dyer, A. W. Black et I. Trancoso. 2015, « Finding Function in Form : Compositional Character Models for Open Vocabulary Word Representation », dans *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, Lisbonne, Portugal, p. 1520–1530.

Lippincott, T., D. Ó Séaghdha et A. Korhonen. 2012, « Learning syntactic verb frames using graphical models », dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, République de Corée, p. 420–429.

List, J.-M. 2014, *Sequence comparison in historical linguistics*, Düsseldorf University Press, Düsseldorf, Allemagne.

List, J.-M., S. J. Greenhill et R. D. Gray. 2017, « The potential of automatic word comparison for historical linguistics », *PLOS ONE*, vol. 12, nº 1, p. 1–18.

Liu, H. 2003, « Unpacking meaning from words: A context-centered approach to computational lexicon design », dans *Modeling and Using Context : Fourth International and Interdisciplinary Conference, Context 2003*, édité par P. Blackburn, C. Ghidini, R. M. Turner et F. Giunchiglia, Springer-Verlag, Stanford, Californie, États-Unis, p. 218–232.

Liu, P.-P., W.-J. Li, N. Lin et X.-S. Li. 2013, « Do Chinese readers follow the National Standard Rules for word segmentation during reading? », *PLoS ONE*, vol. 8, nº 2.

Lopatková, M., Z. Žabokrtský, E. Bejček, K. Skwarska et V. Kettnerová. 2008, *Valenční slovník českých sloves*, Karolinum, Prague, République tchèque.

Lux-Pogodalla, V. et A. Polguère. 2011, « Construction of a French Lexical Network: Methodological Issues », dans *Proceedings of the International Workshop on Lexical Recources (WoLeR 2011)*, Ljubljana, Slovénie.

Macleod, C., R. Grishman, A. Meyers, L. Barrett et R. Reeves. 1998, « Nomlex: A lexicon of nominalizations », dans *Proceedings of the 8th EURALEX International Congress*, Liège, Belgique, p. 488–494.

Magerman, D. M. 1995, « Statistical decision-tree models for parsing », dans *Proceedings of the 33rd annual meeting on Association for Computational Linguistics (ACL 1995)*, Cambridge, Massachusetts, États-Unis, p. 276–283.

Magistry, P. 2012, « Segmentation non supervisée : le cas du mandarin », dans *Actes des Rencontres des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2012)*, Grenoble, France.

Magistry, P. 2013, *Unsupervised Word Segmentation and Wordhood Assessment — The case for Mandarin Chinese*, Thèse de doctorat, Université Denis–Diderot Paris 7, Paris, France.

Magistry, P. et B. Sagot. 2011, « Segmentation et induction de lexiques non-supervisées du mandarin », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France.

Magistry, P. et B. Sagot. 2012, « Unsupervized word segmentation: the case for Mandarin Chinese », dans *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, Jeju, République de Corée.

Magistry, P. et B. Sagot. 2013, « Can MDL Improve Unsupervised Chinese Word Segmentation? », dans *Proceedings of the SIGHAN workshop at IJCNLP 2013*, Nagoya, Japon.

Makkai, A. 1972, *Idiom structure in English*, Janua linguarum : Series maior, Mouton, La Haye, Pays-Bas.

Malouf, R. et F. Ackerman. 2010, « Paradigms: The low entropy conjecture », Paper presented at the Workshop on Morphology and Formal Grammar.

Mann, G. S. et D. Yarowsky. 2001, « Multipath translation lexicon induction via bridge languages », dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2001)*, p. 151–158.

Manning, C. 1995, « Dissociating functor-argument structure from surface phrase structure: the relationship of HPSG Order Domains to LFG », Ms., Carnegie Mellon University.

Manning, C. D. 1993, « Automatic acquisition of a large subcategorization dictionary from corpora », dans *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL 1993)*, Columbus, Ohio, États-Unis, p. 235–242.

Manning, C. D. 2011, « Part-of-speech Tagging from 97% to 100%: Is It Time for Some Linguistics? », dans *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, Tokyo, Japon, p. 171–189.

Manning, C. D. 2015, « Computational linguistics and deep learning », *Computational Linguistics*, vol. 41, nº 4, p. 701–707.

Manning, C. D. et H. Schütze. 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, États-Unis.

Marcus, M. P., M. A. Marcinkiewicz et B. Santorini. 1993, « Building a large annotated corpus of English: The Penn Treebank », *Computational Linguistics*, vol. 19, nº 2, p. 313–330.

Marimon, M., N. Seghezzi et N. Bel. 2007, « An open-source lexicon for Spanish », dans *Sociedad Española para el Procesamiento del Lenguaje Natural, n. 39.*

Marshall, I. 1983, « Choice of grammatical word-class without global syntactic analysis: Tagging words in the lob corpus. », *Computers and the Humanities*, vol. 17, n° 3, p. 139–150.

Marshall, I. 1987, « Tag selection using probabilistic methods », dans *The Computational analysis of English : a corpus-based approach*, édité par R. Garside, G. Sampson et G. Leech, Longman, Londres, Royaume-Uni, p. 42–65.

Martin, R. 1969, « Le trésor de la langue française et la méthode lexicographique », *Langue française*, vol. 2, n° 1, p. 44–55.

Martins, A. F. T., M. B. Almeida et N. A. Smith. 2013, « Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers », dans *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgarie, p. 617–622.

Martins, A. F. T., N. A. Smith, E. P. Xing, P. M. Q. Aguiar et M. A. T. Figueiredo. 2010, « Turbo parsers: Dependency parsing by approximate variational inference », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Cambridge, Massachusetts, États-Unis, p. 34–44.

Matos, R., L. Ferrand, C. Pallier et B. New. 2001, « Une base de données lexicales du français contemporain sur internet », *L'année psychologique*, vol. 101, n° 3, p. 447–462.

Matsuzaki, T., Y. Miyao et J. Tsujii. 2005, « Probabilistic CFG with Latent Annotations », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, États-Unis, p. 75–82.

Matthews, P. H. 1972, *Inflectional Morphology: a Theoretical Study Based on Aspect of Latin Verb Conjugation*, Cambridge University Press, Cambridge, Royaume-Uni.

Matthews, P. H. 1974, *Morphology*, Cambridge University Press, Cambridge, Royaume-Uni.

Matthews, P. H. 1991, *Morphology (2ème édition)*, Cambridge University Press, Cambridge, Royaume-Uni.

Matuszek, C., J. Cabral, M. Witbrock et J. Deoliveira. 2006, « An introduction to the syntax and content of Cyc », dans *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, p. 44–49.

Maurel, D. 2008, « Prolexbase: a multilingual relational lexical database of proper names », dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc, p. 334–338.

Max, A. et G. Wisniewski. 2010, « Mining naturally-occurring corrections and paraphrases from wikipedia's revision history », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

McCarthy, J. J. 2005, « Optimal Paradigms », dans *Paradigms in Phonological Theory*, édité par L. J. Downing, T. A. Hall et R. Raffelsiefen, Oxford University Press, p. 170–210.

McClosky, D., E. Charniak et M. Johnson. 2006, « Reranking and self-training for parser adaptation », dans *Proceedings of COLING-ACL 2006*, Sydney, Australie, p. 337–344.

McDonald, R., K. Crammer et F. Pereira. 2005, « Online large-margin training of dependency parsers », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, États-Unis, p. 91–98.

McGillivray, B. et M. Passarotti. 2009, « The development of the "Index Thomisticus" treebank valency lexicon », dans *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH – SHELT&R 2009)*, Athènes, Grèce.

McWhorter, J. 2001, « The world's simplest grammars are creole grammars », *Linguistic Typology*, vol. 5, p. 125–66.

Meľčuk, I. A. 1974, *Опыт теории лингвистических моделей Смысл ⇔ Текст (Esquisse d'une théorie des modèles linguistiques Sens ⇔ Texte) — Vol 1. Семантика, синтаксис (Sémantique, syntaxe)*, Наука, Moscou, Union des Républiques Socialistes Soviétiques.

Meľčuk, I. A. 1988, *Dependency Syntax: Theory and Practice*, State University Press of New York, New York City, New York, États-Unis.

Meľčuk, I. A. et A. Polguère. 1995, *Introduction à la lexicologie explicative et combinatoire*, Duculot/AUPELF-UREF, Louvain-la-Neuve/Paris.

de Melo, G. et G. Weikum. 2009, « Towards a universal wordnet by learning from combined evidence », dans *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09)*, Hong Kong, Chine, p. 513–522.

Merialdo, B. 1994, « Tagging English Text with a Probabilistic Model », *Computational Linguistics*, vol. 20, nº 2, p. 155–171.

Merrilees, B. 1994, « The Shape of the Medieval Dictionary Entry », *CH Working Papers*, vol. 4, p. 49–60.

Messiant, C., A. Korhonen et T. Poibeau. 2008, « LexSchem: A Large Subcategorization Lexicon for French Verbs », dans *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Maroc.

Meľčuk *et al.*, I. A. 1984, 1988, 1992, 1999, *Dictionnaire explicatif et combinatoire du français contemporain — Recherches lexico-sémantiques, vol. I, II, III, IV*, Presses de l'Université de Montréal, Montréal, Québec, Canada.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado et J. Dean. 2013, « Distributed representations of words and phrases and their compositionality », dans *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)*, Lake Tahoe, Nevada, États-Unis, p. 3111–3119.

Milićević, J. 2009, « Schéma de régime : le pont entre le lexique et la grammaire », *Langages*, vol. 176, nº 4.

Milin, P., D. Filipović Đurđević et F. Moscoso del Prado Martín. 2009, « The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian », *Journal of Memory and Language*, vol. 60, nº 1, p. 50–64.

Miller, G. A. 1995, « WordNet: A Lexical Database for English », *Communications of the ACM*, vol. 38, nº 11, p. 39–41.

Mirroshandel, S. et A. Nasr. 2011, « Active Learning for Dependency Parsing Using Partially Annotated Sentences », dans *Proceedings of the 12th International Conference on Parsing Technologies (IWPT 2011)*, Dublin, Irlande.

Mirroshandel, S. A., A. Nasr et B. Sagot. 2013, « Enforcing Subcategorization Constraints in a Parser Using Sub-parses Recombining », dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, Atlanta, Géorgie, États-Unis.

Mitton, R. 1996, *English Spelling and the computer*, Longman, Londres, Royaume-Uni.

Mitton, R. 2010, « Fifty Years of Spellchecking », *Writing Systems Research*, vol. 2, nº 1, p. 1–7.

Molinero, M. A., B. Sagot et L. Nicolas. 2009a, « Building a morphological and syntactic lexicon by merging various linguistic ressources », dans *Proceedings of NODALIDA 2009*, Odense, Danemark.

Molinero, M. A., B. Sagot et L. Nicolas. 2009b, « A morphological and syntactic wide-coverage lexicon for Spanish: The Leffe », dans *Proceedings of the 7th conference on Recent Advances in Natural Language Processing (RANLP 2009)*, Borovets, Bulgarie.

Monson, C., J. Carbonell, A. Lavie et L. Levin. 2008, « ParaMor: Finding Paradigms across Morphology », dans *Advances in Multilingual and Multimodal Information Retrieval, Revised Selected Papers from the 8th Workshop of the Cross-Language Evaluation Forum (CLEF 2007)*, *Lecture Notes in Computer Science (LNCS)*, vol. 5152, édité par C. Peters, V. Jijkoun, T. Mandl, H. Müller, D. W. Oard, A. Peñas, V. Petras et D. Santos, Springer-Verlag, Budapest, Hongrie, p. 900–907.

Montermini, F. et O. Bonami. 2013, « Stem spaces and predictability in verbal inflection », *Lingue e linguaggio*, vol. 2, p. 171–190.

Moortgat, M. 1988, *Categorial investigations: logical and linguistic aspects of the Lambek calculus*, Groningen-Amsterdam studies in semantics, Foris Publications, Dordrecht, Pays-Bas.

Mouton, C. et G. de Chalendar. 2010, « JAWS: Just Another WordNet Subset », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Müller, T., H. Schmid et H. Schütze. 2013, « Efficient higher-order CRFs for morphological tagging », dans *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing*, Seattle, Washington, États-Unis, p. 322–332.

Müller, T. et H. Schütze. 2015, « Robust morphological tagging with word representations », dans *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2015)*, Denver, Colorado, États-Unis, p. 526–536.

Nakamura, T. 2006, *Lexique et grammaire des interrogatives partielles en français : étude des verbes à une complétive directe*, Thèse de doctorat, Université de Marne-la-Vallée.

Nasr, A., F. Béchet, J.-F. Rey, B. Favre et L. R. J. 2011, « MACAON: An NLP tool suite for processing word lattices », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, Portland, Oregon, États-Unis.

Nasr, A., F. Béchet et A. Volanschi. 2004, « Tagging with Hidden Markov Models using ambiguous tags », dans *Proceedings of the 20th International Conference on Computational Linguistics (CoLing 2004)*, Genève, Suisse.

Nastase, V. 2008, « Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, Honolulu, Hawaii, États-Unis, p. 763–772.

Navigli, R. et S. P. Ponzetto. 2010, « BabelNet: Building a very large multilingual semantic network », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède, p. 216–225.

Navigli, R. et S. P. Ponzetto. 2012, « BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network », *Artificial Intelligence*, vol. 193, p. 217–250.

New, B. 2006, « Lexique 3 : Une nouvelle base de données lexicales », dans *Actes de la 13ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique.

Nichols, J. 2009, « Linguistic complexity : a comprehensive definition and survey », dans *Language Complexity as an Evolving Variable*, édité par G. Sampson, D. Gil et P. Trudgill, Oxford, Royaume-Uni, p. 64–79.

Nicolas, L., J. Farré et É. Villemonte de la Clergerie. 2007, « Mining parsing results for lexical corrections », dans *Proceedings of the 3rd Language and Technology Conference (LTC 2007)*, Poznań, Pologne.

Nicolas, L., B. Sagot, M. A. Molinero, J. Farré et É. Villemonte de La Clergerie. 2008a, « Computer aided correction and extension of a syntactic wide-coverage lexicon », dans *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, Royaume-Uni, p. 633–640.

Nicolas, L., B. Sagot, M. A. Molinero, J. Farré et É. Villemonte de La Clergerie. 2008b, « Extensión y corrección semi-automática de léxicos morfo-sintáctico », dans *Proceedings of the 24th Edition of the Conference of the Spanish Society for Natural Language Processing (SEPLN 2008)*, Madrid, Espagne.

Nivre, J., Ž. Agić, L. Ahrenberg, M. J. Aranzabe, M. Asahara, A. Atutxa, M. Ballesteros, J. Bauer, K. Bengoetxea, Y. Berzak, R. A. Bhat, C. Bosco, G. Bouma, S. Bowman, G. Cebiroğlu Eryiğit, G. G. A. Celano, Ç. Çöltekin, M. Connor, M.-C. de Marneffe, A. Diaz de Ilarraza, K. Dobrovoljc, T. Dozat, K. Droganova, T. Erjavec, R. Farkas, J. Foster, D. Galbraith, S. Garza, F. Ginter, I. Goenaga, K. Gojenola, M. Gokirmak, Y. Goldberg, X. Gómez Guinovart, B. Gonzáles Saavedra, N. Grūzītis, B. Guillaume, J. Hajič, D. Haug, B. Hladká, R. Ion, E. Irimia, A. Johannsen, H. Kaşıkara, H. Kanayama, J. Kanerva, B. Katz, J. Kenney, S. Krek, V. Laippala, L. Lam, A. Lenci, N. Ljubešić, O. Lyashevskaya, T. Lynn, A. Makazhanov, C. Manning, C. Mărănduc, D. Mareček, H. Martínez Alonso, J. Mašek, Y. Matsumoto, R. McDonald, A. Missilä, V. Mititelu, Y. Miyao, S. Montemagni, K. S. Mori, S. Mori, K. Muischnek, N. Mustafina, K. Müürisep, V. Nikolaev, H. Nurmi, P. Osenova, L. Øvrelid, E. Pascual, M. Passarotti, C.-A. Perez, S. Petrov, J. Piitulainen, B. Plank, M. Popel, L. Pretkalniņa, P. Prokopidis, T. Puolakainen, S. Pyysalo, L. Ramasamy, L. Rituma, R. Rosa, S. Saleh, B. Saulīte, S. Schuster, W. Seeker, M. Seraji, L. Shakurova, M. Shen, N. Silveira, M. Simi, R. Simionescu, K. Simkó, K. Simov, A. Smith, C. Spadine, A. Suhr, U. Sulubacak, Z. Szántó, T. Tanaka, R. Tsarfaty, F. Tyers, S. Uematsu, L. Uria, G. van Noord, V. Varga, V. Vincze, J. X. Wang, J. N. Washington, Z. Žabokrtský, D. Zeman et H. Zhu. 2016, « Universal dependencies 1.3 », URL `http://hdl.handle.net/11234/1-1699`, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Nivre, J., J. Hall, S. Kübler, R. McDonald, J. Nilsson, S. Riedel et D. Yuret. 2007a, « The CoNLL 2007 shared task on dependency parsing », dans *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, Prague, République tchèque, p. 915–932.

Nivre, J., J. Hall et J. Nilsson. 2006, « Maltparser: A data-driven parser-generator for dependency parsing », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Nivre, J., J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov et E. Marsi. 2007b, « Maltparser: A language-independent system for data-driven dependency parsing », *Natural Language Engineering*, vol. 13, nº 2, p. 95–135.

van Noord, G. 2004, « Error mining for wide-coverage grammar engineering », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelone, Espagne.

Oflazer, K. 1996, « Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction », *Computational Linguistics*, vol. 22, nº 1, p. 73–89.

Oliver, A., I. Castellón et L. Màrquez. 2003, « Use of Internet for augmenting coverage in a lexical acquisition system from raw corpora: application to Russian », dans *Proceedings of the RANLP'03 International Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL'03)*, Borovets, Bulgarie.

Oliver, A. et M. Tadić. 2004, « Enlarging the Croatian morphological lexicon by automatic lexical acquisition from raw corpora », dans *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, p. 1259–1262.

Orav, H. et K. Vider. 2004, « Concerning the difference between a conception and its application in the case of the estonian wordnet », dans *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC 2004)*, Brno, République tchèque, p. 285–290.

Packard, J. L. 2000, *The Morphology of Chinese A Linguistic and Cognitive Approach*, Cambridge University Press, Cambridge, Royaume-Uni.

Palmer, M., D. Gildea et P. Kingsbury. 2005, « The proposition bank: An annotated corpus of semantic roles », *Computational Linguistics*, vol. 31, nº 1, p. 71–106.

Panevová, J. 1994, « Valency Frames and the Meaning of the Sentence », dans *The Prague School of Structural and Functional Linguistics*, édité par P. A. Luelsdorff, John Benjamins, Amsterdam, Pays-Bas, p. 223–243.

Park, Y. A. et R. Levy. 2011, « Automated whole sentence grammar correction using a noisy channel model », dans *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL/HLT 2011)*, Portland, Oregon, États-Unis, p. 934–944.

Paroubek, P., É. Villemonte de la Clergerie, S. Loiseau, A. Vilnat et G. Francopoulo. 2009, « The PASSAGE syntactic representation », dans *Proceedings of the 7th International Workshop on Treebanks and Linguistic Theories (TLT7)*, Groningue, Pays-Bas.

Paroubek, P., I. Robba, A. Vilnat et C. Ayache. 2006, « Data, Annotations and Measures in EASy, the Evaluation Campaign for Parsers of French », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Passarotti, M. C. 2007, « Verso il Lessico Tomistico Biculturale. La treebank dell Index Thomisticus », dans *Intrecci testuali, articolazioni linguistiche, composizioni logiche. Proceedings of the XIII Congresso Nazionale della Società di Filosofia del Linguaggio*, Viterbo, Italie.

Paul, H. 1880, *Prinzipien der Sprachgeschichte*, Max Niemeyer, Halle, Empire Allemand.

Paumier, S. 2003, *De la reconnaissance de formes linguistiques à l'analyse syntaxique*, Thèse de doctorat, Univerité de Marne-la-Vallée.

Pellegrino, F., C. Coupé, et E. Marsico. 2011, « A cross-language perspective on speech information rate », *Language*, vol. 87, nº 3, p. 539–558.

Pellegrino, F., C. Coupé et E. Marsico. 2007, « An information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax », dans *Proceedings of the Annual Meeting of the Linguistic Society of America*, Anaheim, Californie, États-Unis.

Perera, P. et R. Witte. 2005, « A self-learning context-aware lemmatizer for German », dans *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, Colombie Britannique, Canada, p. 636–643.

Perlmutter, D. et P. Postal. 1983, *Studies in Relational Grammar 1*, University of Chicago Press, Chicago, Illinois, États-Unis.

Peters, P. S., Jr. et R. W. Ritchie. 1973, « On the generative power of transformational grammars », *Information Sciences*, vol. 6, p. 49–83.

Peterson, J. 2008, *Kharia: A South Munda language*, Habilitation à diriger des recherches, Universität Osnabrück.

Petrov, S. 2010, « Products of random latent variable grammars », dans *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT 2010)*, Los Angeles, Californie, États-Unis, p. 19–27.

Petrov, S., L. Barrett, R. Thibaux et D. Klein. 2006, « Learning accurate, compact, and interpretable tree annotation », dans *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (CoLing/ACL 2006)*, Sydney, Australie, p. 433–440.

Petrov, S., D. Das et R. McDonald. 2012, « A universal part-of-speech tagset », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, p. 2089–2096.

Petrov, S. et D. Klein. 2007, « Improved inference for unlexicalized parsing », dans *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL/HLT 2007)*, Rochester, New York, États-Unis, p. 404–411.

Petrov, S. et D. Klein. 2008, « Parsing German with Latent Variable Grammars », dans *Proceedings of the ACL 2008 Workshop on Parsing German*, Columbus, Ohio, États-Unis.

Petrov, S. et R. McDonald. 2012, « Overview of the 2012 Shared Task on Parsing the Web », dans *Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL 2012), a NAACL/HLT 2012 workshop*, Montréal, Québec, Canada.

Pianta, E., L. Bentivogli et C. Girardi. 2004, « Fighting arbitrariness in wordnet-like lexical databases – a natural language motivated remedy », dans *Proceedings of the 1st International Conference of the Global WordNet Association (GWC 2002)*, Mysore, Inde.

Pirelli, V. et M. Battista. 2000, « The paradigmatic dimension of stem allomorphy in italian verb inflection », *Rivista di linguistica*, vol. 12, n° 2, p. 307–380.

Plaehn, O. 2005, « Computing the most probable parse for a discontinuous phrase structure grammar », dans *New Developments in Parsing Technology*, *Text, Speech and Language Technology*, vol. 23, édité par H. Bunt, J. Carroll et G. Satta, Springer Netherlands, Dordrecht, Pays-Bas, p. 91–106.

Plank, B., A. Søgaard et Y. Goldberg. 2016, « Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss », dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Allemagne.

Pollard, C. et I. A. Sag. 1987, *Information-based Syntax and Semantics: Vol. 1: Fundamentals*, CSLI Lecture Notes 13, Center for the Study of Language and Information, Stanford, Californie, États-Unis.

Ponzetto, S. P. et R. Navigli. 2009, « Large-scale taxonomy mapping for restructuring and integrating wikipedia », dans *Proceedings of the 21st international jont conference on Artifical intelligence (IJCAI'09)*, Pasadena, Californie, États-Unis, p. 2083–2088.

Pradet, Q., G. de Chalendar et J. Desormeaux Baguenier. 2014a, « WoNeF, an improved, expanded and evaluated automatic French translation of WordNet », dans *Proceedings of the 7th Global Wordnet Conference (GWC 2014)*, p. 32–39.

Pradet, Q., L. Danlos et G. D. Chalendar. 2014b, « Adapting VerbNet to French using existing resources », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Moscoso del Prado Martín, F. 2011, « The Mirage of morphological complexity », dans *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Boston, Massachusetts, États-Unis, p. 3524–3529.

Moscoso del Prado Martín, F., A. Kostić et R. H. Baayen. 2004, « Putting the bits together: An information theoretical perspective on morphological processing », *Cognition*, vol. 94, p. 1–18.

Preiss, J., T. Briscoe et A. Korhonen. 2007, « A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora », dans *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, République tchèque.

Prokić, J. et M. Cysouw. 2013, « Combining Regular Sound Correspondences and Geographic Spread », *Language Dynamics and Change*, vol. 3, p. 147–168.

Przepiórkowski, A. 2009, « Towards the automatic acquisition of a valence dictionary for polish », dans *Aspects of Natural Language Processing, Lecture Notes in Computer Science (LNCS)*, vol. 5070, édité par M. Marciniak et A. Mykowiecka, Springer-Verlag, Berlin, Allemagne, p. 191–210.

Przepiórkowski, A., E. Hajnicz, A. Patejuk, M. Woliński, F. Skwarski et M. Świdziński. 2014, « Walenty: Towards a comprehensive valence dictionary of polish », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

QasemiZadeh, B. et S. Rahimi. 2006, « Persian in MULTEXT-East Framework », dans *Advances in Natural Language Processing, 5th International Conference on NLP (FinTAL 2006)*, Turku, Finlande, p. 541–551.

Radeau, M., P. Mousty et A. Content. 1990, « Brulex. une base de données lexicales informatisée pour le français écrit et parlé », *L'année psychologique*, vol. 90, nº 4, p. 551–566.

Rapp, R. 1999, « Automatic identification of word translations from unrelated English and German corpora », dans *Proceedings of ACL 1999*, p. 519–526.

Ratnaparkhi, A. 1996, « A Maximum Entropy Model for Part-Of-Speech Tagging », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 1996)*, Philadelphie, Pennsylvanie, États-Unis, p. 133–142.

Rauzy, S. et P. Blache. 2007, « Un lexique syntaxique des verbes du français : VfrLPL », cahier de recherche RAU-3055, Laboratoire Parole et Langage.

Rauzy, S. et P. Blache. 2009, « Un point sur les outils du lpl pour l'analyse syntaxique du français », dans *Journée ATALA « Quels analyseurs syntaxiques pour le français ? » (organisée conjointement à la conférence IWPT 2009)*, édité par É. Villemonte de la Clergerie et P. Paroubek, Paris, France.

Ravi, S. et K. Knight. 2009, « Minimized models for unsupervised part-of-speech tagging », dans *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP'09)*, Singapour, Singapour, p. 504–512.

Reape, M. 1993, *A Formal Theory of Word Order: A Case Study in West Germanic*, University of Edinburgh, Édimbourg, Royaume-Uni.

Reiter, N., M. Hartung et A. Frank. 2008, « A Resource-Poor Approach for Linking Ontology Classes to Wikipedia Articles », dans *Semantics in Text Processing. STEP 2008 Conference Proceedings*, *Research in Computational Semantics*, vol. 1, édité par J. Bos et R. Delmonte, College Publications, p. 381–387.

Resnik, P. 1992, « Probabilistic tree-adjoining grammar as a framework for statistical natural language processing », dans *Proceedings of the 14th International Conference on Computational Linguistics (CoLing 1992)*, Nantes, France, p. 418–424.

Resnik, P. et D. Yarowsky. 1997, « A perspective on word sense disambiguation methods and their evaluation », dans *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How ?*, Washington, D.C., États-Unis, p. 79–86.

Richardson, S. D., W. B. Dolan et L. Vanderwende. 1998, « Mindnet: Acquiring and structuring semantic information from text », dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montréal, Québec, Canada, p. 1098–1102.

Riezler, S., T. H. King, R. M. Kaplan, R. Crouch, J. T. Maxwell, III et M. Johnson. 2002, « Parsing the Wall Street Journal using a Lexical-Functional Grammar and discriminative estimation techniques », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphie, Pennsylvanie, États-Unis, p. 271–278.

Rissanen, J. 1984, « Universal coding, information, prediction, and estimation », *IEEE Transactions on Information Theory*, vol. 30, nº 4, p. 629–636.

Roland, D. et D. Jurafsky. 1998, « How verb subcategorization frequencies are affected by corpus choice », dans *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL '98)*, Montréal, Québec, Canada, p. 1122–1128.

Role, F., M. F. Gavilanes et É. Villemonte de la Clergerie. 2007, « Large-scale knowledge acquisition from botanical texts », dans *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems (NLDB 2007)*, Paris, France.

Romary, L., S. Salmon-Alt et G. Francopoulo. 2004, « Standards going concrete: from lmf to morphalou », dans *Proceedings of the Coling 2004 Workshop on Electronic Dictionaries*, Genève, Suisse.

Rudnicka, E., M. Maziarz, M. Piasecki et S. Szpakowicz. 2012, « A strategy of mapping Polish Wordnet onto Princeton Wordnet », dans *Proceedings of the 24th International Conference on Computational Linguistics (CoLing 2012)*, Bombay, Inde, p. 1039–1048.

Ruiz-Casado, M., E. Alfonseca et P. Castells. 2005, « Automatic assignment of wikipedia encyclopedic entries to wordnet synsets », dans *Proceedings of the Advances in Web Intelligence Third International Atlantic Web Intelligence Conference (AWIC 2005)*, Łódź, Pologne.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck et C. R. Johnson. 2005, « FrameNet II: Extended Theory and Practice », cahier de recherche, ICSI. Version mise à jour disponible à l'adresse http ://FrameNet.icsi.berkeley.edu/book/book.html.

Ryder, R. J. et G. K. Nicholls. 2011, « Missing data in a stochastic Dollo model for cognate data, and its application to the dating of Proto-Indo-European », *Journal of the Royal Statistical Society Series C (Applied Statistics)*, vol. 60, n° 1, p. 71–92.

Sablayrolles, J. 1997, « Néologismes : Une typologie des typologies », *Cahier du CIEL*, vol. 1996-1997, p. 11–48.

Sagot, B. 2005a, « Automatic acquisition of a Slovak lexicon from a raw corpus », dans *Lecture Notes in Artificial Intelligence 3658 (Proceedings of the Text, Speech and Dialogue 2005 conference)*, Springer-Verlag, Karlovy Vary, République tchèque, p. 156–163.

Sagot, B. 2005b, « Les Méta-RCG : description et mise en oeuvre », dans *Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France, p. 493–498.

Sagot, B. 2006, *Analyse automatique du français : lexiques, formalismes, analyseurs*, Thèse de doctorat, Université Paris 7.

Sagot, B. 2007, « Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish », dans *Proceedings of the 2nd Language and Technology Conference (LTC 2005)*, Poznań, Pologne, p. 423–427.

Sagot, B. 2010, « The Le*fff*, a freely available, accurate and large-coverage lexicon for french », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Sagot, B. 2013a, « Comparing Complexity Measures », dans *Computational approaches to morphological complexity*, Surrey Morphology Group, Paris, France.

Sagot, B. 2013b, « Construction de ressources lexicales pour le traitement automatique des langues », dans *Ressources Lexicales – Contenu, construction, utilisation, évaluation, Lingvisticæ Investigationes Supplementa*, vol. 30, édité par N. Gala et M. Zock, John Benjamins, p. 217–254.

Sagot, B. 2014, « DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Sagot, B. 2016a, « External Lexical Information for Multilingual Part-of-Speech Tagging », Research Report RR-8924, Inria.

Sagot, B. 2016b, « Multilingual part-of-speech tagging with MElt », dans *Actes de la 23ème Conférence sur le Traitement Automatique des Langues Naturelles*, Paris, France.

Sagot, B. 2017a, « Construction automatique d'une base de données étymologiques à partir du wiktionary », dans *Actes de la 24ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans, France.

Sagot, B. 2017b, « Extracting an Etymological Database from Wiktionary », dans *Electronic Lexicography in the 21st century (eLex 2017)*, Leyde, Pays-Bas, p. 716–728.

Sagot, B. 2017c, « Représentation de l'information sémantique lexicale : le modèle wordnet et son application au français », *Revue Française de Linguistique Appliquée*, vol. 22, p. 131–146.

Sagot, B. 2018a, « A new PIE root **h₁er* '(to be) dark red, dusk red' : drawing the line between inherited and borrowed words for 'red(ish)', 'pea', 'ore', 'dusk' and 'love' in daughter languages », dans *International Colloquium on Loanwords and Substrata in Indo-European languages*, Limoges, France.

Sagot, B. 2018b, « A multilingual collection of conll-u-compatible morphological lexicons », dans *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japon.

Sagot, B. et P. Boullier. 2004, « Les RCG comme formalisme grammatical pour la linguistique », dans *Actes de la 11ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*, Fès, Maroc, p. 403–412.

Sagot, B. et P. Boullier. 2005a, « From raw corpus to word lattices: robust pre-parsing processing », dans *Proceedings of the 2nd Language and Technology Conference (LTC 2005)*, Poznań, Pologne, p. 348–351.

Sagot, B. et P. Boullier. 2005b, « From raw corpus to word lattices: robust pre-parsing processing with SxPipe », *Archives of Control Sciences (Special Issue on Language and Technology)*, vol. 15, n° 4, p. 653–662.

Sagot, B. et P. Boullier. 2006, « Deep non-probabilistic parsing of large corpora », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Gênes, Italie.

Sagot, B. et P. Boullier. 2008, « SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts », *Traitement Automatique des Langues*, vol. 49, n° 2, p. 155–188.

Sagot, B., L. Clément, É. Villemonte de La Clergerie et P. Boullier. 2006, « The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use », dans *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, Lisbonne, Portugal.

Sagot, B. et L. Danlos. 2007, « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire : Constructions impersonnelles et expressions verbales figées », dans *Actes de la conférence sur la Description linguistique pour le traitement automatique du français*, Cahiers du CENTAL, Katolieke Universiteit Leuven, Louvain, Belgique.

Sagot, B. et L. Danlos. 2008, « Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français », dans *Actes du colloque Lexicographie et informatique : bilan et perspectives*, Nancy, France.

Sagot, B. et L. Danlos. 2009, « Constructions pronominales dans dicovalence et le lexique-grammaire — intégration dans le lefff », *Linguisticæ Investigationes*, vol. 32, n° 2.

Sagot, B. et L. Danlos. 2010, « Verbes de citation et tables du lexique-grammaire », dans *Actes du 29e Colloque sur le Lexique et la Grammaire*, Belgrade, Serbie.

Sagot, B. et L. Danlos. 2012, « Merging syntactic lexica: the case for french verbs », dans *Proceedings of the LREC 2012 Workshop on Merging Language Resources*, Istanbul, Turquie.

Sagot, B., L. Danlos et M. Colinet. 2014, « Sous-catégorisation en *pour* et syntaxe lexicale », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France.

Sagot, B., L. Danlos et R. Stern. 2010, « A lexicon of french quotation verbs for automatic quotation extraction », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Sagot, B. et D. Fišer. 2008, « Building a free French wordnet from multilingual resources », dans *Proceedings of Ontolex 2008*, Marrakech, Maroc.

Sagot, B. et D. Fišer. 2011, « Extending wordnets by learning from multiple resources », dans *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, Poznań, Pologne.

Sagot, B. et D. Fišer. 2012a, « Automatic Extension of WOLF », dans *Proceedings of the 6th International Global Wordnet Conference (GWC2012)*, Matsue, Japon.

Sagot, B. et D. Fišer. 2012b, « Cleaning noisy wordnets », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie.

Sagot, B. et D. Fišer. 2014, « Classification-based extension of wordnets from heterogeneous resources », dans *Human Language Technology Challenges for Computer Science*

and Linguistics, Lecture Notes in Computer Science, Springer-Verlag, Poznań, Pologne, p. 396–407.

Sagot, B. et K. Fort. 2007, « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire – adverbes en *-ment.* », dans *Actes du 26e Colloque sur le Lexique et la Grammaire*, Bonifacio, France.

Sagot, B. et K. Fort. 2009, « Description et analyse des verbes désadjectivaux et dénominaux en *-ifier* et *-iser* », dans *Proceedings of the 28th Lexis and Grammar Conference*, Bergen, Norvège.

Sagot, B., K. Fort, G. Adda, J. Mariani et B. Lang. 2011a, « Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France.

Sagot, B., K. Fort et F. Venant. 2008, « Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes », dans *Proceedings of the 27th Conference on Lexis and Grammar*, L'Aquila, Italie.

Sagot, B., K. Fort et F. Venant. 2009a, « Extending the adverbial coverage of a french wordnet », dans *Proceedings of the NODALIDA 2009 workshop on WordNets and other Lexical Semantic Resources*, Odense, Danemark.

Sagot, B., K. Fort et F. Venant. 2009b, « Extension et couplage de ressources syntaxiques et sémantiques sur les adverbes », *Linguisticæ Investigationes*, vol. 32, nº 2.

Sagot, B. et K. Gábor. 2014, « Détection et correction automatique d'entités nommées dans des corpus OCRisés », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France.

Sagot, B. et H. Martínez Alonso. 2017, « Improving neural tagging with lexical information », dans *Proceedings of the 15th International Conference on Parsing Technologies (IWPT 2017)*, Pise, Italie, p. 25–31.

Sagot, B., D. Nouvel, V. Mouilleron et M. Baranes. 2013, « Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel », dans *Actes de la 20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, France, p. 407–420.

Sagot, B., M. Richard et R. Stern. 2012, « Annotation référentielle du Corpus Arboré de Paris 7 en entités nommées », dans *Actes de la 19ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2012)*, Grenoble, France.

Sagot, B. et G. Satta. 2010, « Optimal rank reduction for Linear Context-Free Rewriting Systems with Fan-Out Two », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède.

Sagot, B. et R. Stern. 2012, « Aleda, a free large-scale entity database for French », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, p. 1273–1276.

Sagot, B. et E. Tolone. 2009a, « Exploitation des tables du Lexique-Grammaire pour l'analyse syntaxique automatique », *Arena Romanistica*, vol. 4 (Proceedings of the 28th Conference on Lexis and Grammar), p. 302–312.

Sagot, B. et E. Tolone. 2009b, « Intégrer les tables du Lexique-Grammaire à un analyseur syntaxique robuste à grande échelle », dans *Actes de la 16ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France.

Sagot, B. et É. Villemonte de La Clergerie. 2006, « Error mining in parsing results », dans *Proceedings of ACL/COLING 2006*, Sydney, Australie, p. 329–336.

Sagot, B. et É. Villemonte de La Clergerie. 2008, « Fouille d'erreurs sur des sorties d'analyseurs syntaxiques », *Traitement Automatique des Langues*, vol. 49, n° 1.

Sagot, B. et G. Walther. 2010a, « Développement de ressources pour le persan : lexique morphologique et chaîne de traitements de surface », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Sagot, B. et G. Walther. 2010b, « A morphological lexicon for the Persian language », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Sagot, B. et G. Walther. 2011, « Non-canonical inflection : data, formalisation and complexity measures. », dans *Proceedings of the workshop on Systems and Frameworks in Computational Morphology (SFCM 2011)*, vol. 100, édité par C. Mahlow et M. Piotrowski, Springer-Verlag, Zurich, Suisse, p. 23–45.

Sagot, B. et G. Walther. 2013, « Implementing a formal model of inflectional morphology », dans *Proceedings of the 3rd International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2013)*, *Communications in Computer and Information Science (CCIS)*, vol. 380, Springer-Verlag, Berlin, Allemagne, p. 115–134.

Sagot, B., G. Walther, P. Faghiri et P. Samvelian. 2011b, « A new morphological lexicon and a POS tagger for the Persian Language », dans *Proceedings of the International Conference in Iranian Linguistics (ICIL 2011)*, Uppsala, Suède.

Sagot, B., G. Walther, P. Faghiri et P. Samvelian. 2011c, « Développement de ressources pour le persan : PerLex 2 et MElt_{fa} », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France.

Sajous, F., N. Hathout et B. Calderone. 2013, « GLÀFF, un Gros Lexique À tout Faire du Français », dans *Actes de la 20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, France, p. 285–298.

Samvelian, P., L. Danlos et B. Sagot. 2011, « On the predictability of light verbs », dans *Proceedings of the 30th International Conference on Lexis and Grammar*, Nicosie, Chypre.

Samvelian, P., P. Faghiri et S. E. Ayari. 2014, « Extending the coverage of a mwe database for persian cps exploiting valency alternations », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande, p. 4023–4026.

Sapir, E. 1921, *Language: An Introduction to the Study of Speech*, Harcourt Brace, New York City, New York, États-Unis.

de Saussure, F. 1916, *Cours de linguistique générale*, Payot & Rivages (édition critique publiée en 1997), Paris, France.

Scherrer, Y. et B. Sagot. 2013a, « Étiquetage morphosyntaxique de langues non dotées à partir de ressources pour une langue étymologiquement proche », dans *Actes du workshop TALARE de la conférence TALN 2013*, Les Sables d'Olonne, France.

Scherrer, Y. et B. Sagot. 2013b, « Lexicon induction and part-of-speech tagging of non-resourced languages without any bilingual resources », dans *Proceedings of the RANLP 2013 workshop on Adaptation of language resources and tools for closely related languages and language variants*, Hissar, Bulgarie.

Scherrer, Y. et B. Sagot. 2014, « A language-independent and fully unsupervised approach to lexicon induction and part-of-speech tagging for closely related languages », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Schmid, H. 1994, « Probabilistic part-of-speech tagging using decision trees », dans *Proceedings of International Conference on New Methods in Language Processing*, Manchester, Royaume-Uni.

Schmid, H., A. Fitschen et U. Heid. 2004, « SMOR: A german computational morphology covering derivation, composition, and inflection », dans *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbonne, Portugal, p. 1263–1266.

Schreuder, R. et R. H. Baayen. 1997, « How complex simplex words can be », *Journal of Memory and Language*, vol. 36, p. 118–139.

Schuster, S., É. Villemonte de La Clergerie, M. Candito, B. Sagot, C. D. Manning et D. Seddah. 2017, « Paris and Stanford at EPE 2017 : Downstream Evaluation of Graph-based Dependency Representations », dans *Proceedings of the 1st Shared Task on Extrinsic Parser Evaluation (EPE 2017)*, Pise, Italie, p. 47–59.

Seddah, D., M. Candito et B. Crabbé. 2009, « Cross parser evaluation and tagset variation: A French Treebank study », dans *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France, p. 150–161.

Seddah, D., M. Candito, B. Crabbé et E. H. Anguiano. 2012a, « Ubiquitous usage of a broad coverage french corpus: Processing the est republicain corpus », dans *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turquie, p. 3249–3254.

Seddah, D., G. Chrupała, O. Cetinoglu, J. van Genabith et M. Candito. 2010a, « Lemmatization and lexicalized statistical parsing of morphologically-rich languages: the case of french », dans *Proceedings of the NAACL/HLT 2010 Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010)*, Los Angeles, Californie, États-Unis, p. 85–93.

Seddah, D., S. Kübler et R. Tsarfaty. 2014, « Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages », dans *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, Dublin, Irlande, p. 103–109.

Seddah, D., J. Le Roux et B. Sagot. 2011, « Data Driven Lemmatization for Statistical Constituent Parsing of Italian », dans *Proceedings of EVALITA 2011*, Springer-Verlag, Rome, Italie.

Seddah, D., J. Le Roux et B. Sagot. 2013a, « Data driven lemmatization and parsing of italian », dans *Evaluation of Natural Language and Speech Tools for Italian (Revised Selected Papers)*, *Lecture Notes in Computer Science*, vol. 7689, édité par B. Magnini, F. Cutugno, M. Falcone et E. Pianta, Springer-Verlag, Rome, Italie, p. 249–256.

Seddah, D. et B. Sagot. 2006a, « Modeling and Analysis of Elliptic Coordination by Dynamic Exploitation of Derivation Forests in LTAG Parsing », dans *Proceedings of TAG+8 : The Eighth International Workshop on Tree Adjoining Grammar and Related Formalisms*, Sydney, Australie.

Seddah, D. et B. Sagot. 2006b, « Modélisation et analyse des coordinations elliptiques par l'exploitation dynamique des forêts de dérivation », dans *Actes de la 13ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique, p. 609–618.

Seddah, D., B. Sagot et M. Candito. 2012b, « The Alpage Architecture at the SANCL 2012 Shared Task: Robust Pre-Processing and Lexical Bridging for User-Generated Content Parsing », dans *Proceedings of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL 2012), a NAACL/HLT 2012 workshop*, Montréal, Québec, Canada.

Seddah, D., B. Sagot, M. Candito, V. Mouilleron et V. Combet. 2012c, « Building a treebank of noisy user-generated content: The French Social Media Bank », dans *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories (TLT11)*, Lisbonne, Portugal.

Seddah, D., B. Sagot, M. Candito, V. Mouilleron et V. Combet. 2012d, « The French Social Media Bank: a Treebank of Noisy User Generated Content », dans *Proceedings of the 24th International Conference on Computational Linguistics (CoLing 2012)*, Bombay, Inde.

Seddah, D., B. Sagot et L. Danlos. 2010b, « Control Verbs, Argument Cluster Coordination and MCTAG », dans *Proceedings of the 10th International Conference on Tree Adjoining Grammars and Related Formalisms (TAG+10)*, New Haven, Connecticut, États-Unis.

Seddah, D., R. Tsarfaty, S. Kübler, M. Candito, J. D. Choi, R. Farkas, J. Foster, I. Goenaga, K. Gojenola Galletebeitia, Y. Goldberg, S. Green, N. Habash, M. Kuhlmann, W. Maier, J. Nivre, A. Przepiórkowski, R. Roth, W. Seeker, Y. Versley, V. Vincze, M. Woliński, A. Wróblewska et É. Villemonte de La Clergerie. 2013b, « Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages », dans *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, Seattle, Washington, États-Unis, p. 146–182.

Sennrich, R. et B. Kunz. 2014, « Zmorge: A german morphological lexicon extracted from wiktionary », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Sgall, P., E. Hajicová et J. Panevová. 1986, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, Springer-Verlag, Berlin, République Fédérale d'Allemagne.

Shannon, C. E. 1948, « A mathematical theory of communication », *Bell System Technical Journal*, vol. 27, p. 379–423,623–656.

Shieber, S. M. 1986, *An introduction to unification-based approaches to grammar*, n° 4 dans CSLI lecture notes, Center for the Study of Language and Information, Stanford, Californie, États-Unis.

Shieber, S. M. 1987, « Evidence against the context-freeness of natural language », dans *The Formal Complexity of Natural Language*, *Studies in Linguistics and Philosophy*, vol. 33, édité par W. Savitch, E. Bach, W. Marsh et G. Safran-Naveh, Pays-Bas : Springer, p. 320–334.

Shosted, R. K. 2006, « Correlating complexity: A typological approach », *Linguistic Typology*, vol. 10, p. 1–40.

Sigogne, A. et M. Constant. 2012, « Using subcategorization frames to improve French probabilistic parsing », dans *Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012)*, Vienne, Autriche, p. 223–227.

Silberztein, M. 1990, « Le dictionnaire électronique des mots composés », *Langue française*, vol. 87, n° 1, p. 71–83.

Silverstein, M. 1976, « Hierarchy of Features and Ergativity », dans *Grammatical Categories in Australian Languages*, édité par R. M. Dixon, Australian Institute of Aboriginal Studies, Canberra, Australie.

Smith, G. 2003, « A brief introduction to the TIGER treebank, version 1 », cahier de recherche, Universität Potsdam.

Smith, N. et J. Eisner. 2005, « Contrastive estimation: Training log-linear models on unlabeled data », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, États-Unis, p. 354–362.

Snover, M. G. et M. R. Brent. 2001, « A Bayesian model for morpheme and paradigm identification », dans *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, Toulouse, France, p. 490–498.

Solomonoff, R. 1960, « A preliminary report on a general theory of inductive inference », cahier de recherche V-131, Zator Co., Cambridge, Massachusetts, États-Unis.

Solomonoff, R. 1964, « A formal theory of inductive inference », *Information and Control*, vol. 7, p. 1–22, 224–254.

Sornlertlamvanich, V. 2010, « Asian WordNet: Development and service in collaborative approach », dans *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Bombay, Inde.

Steele, S. 1995, « Towards a theory of morphological information », *Language*, vol. 71, nº 2, p. 260–309.

Stern, R. 2015, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Thèse de doctorat, Université Denis–Diderot Paris 7.

Stern, R. et B. Sagot. 2010a, « Détection et résolution d'entités nommées dans des dépêches d'agence », dans *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, Montréal, Québec, Canada.

Stern, R. et B. Sagot. 2010b, « Resources for named entity recognition and resolution in news wires », dans *Entity 2010 Workshop at LREC 2010*, La Valette, Malte.

Stern, R. et B. Sagot. 2012, « Population of a knowledge base for news metadata from unstructured text and web data », dans *Proceedings of the Knowledge Extraction Workshop at NAACL-HLT 2012 (AKBC-WEKEX 2012)*, Montréal, Québec, Canada, p. –.

Stern, R., B. Sagot et F. Béchet. 2012, « A joint named entity recognition and entity linking system », dans *Proceedings of the EACL 2012 Workshop on Innovative hybrid approaches to the processing of textual data*, Avignon, France.

Straka, M., J. Hajič et J. Straková. 2016, « UDPipe : trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing », dans *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovénie.

Strnadová, J. et B. Sagot. 2011, « Construction d'un lexique des adjectifs dénominaux », dans *Actes de la 18ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier, France, p. 69–74.

Stroppa, N. et F. Yvon. 2005, « An analogical learner for morphological analysis », dans *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*, Ann Arbor, Michigan, États-Unis, p. 120–127.

Stroppa, N. et F. Yvon. 2006, « Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie », *Traitement Automatique des Langues*, vol. 47, nº 1, p. 33–59.

Stump, G. et R. Finkel. 2013, *Morphological Typology: From Word to Paradigm*, Cambridge Studies in Linguistics, Cambridge University Press, Cambridge, Royaume-Uni.

Stump, G. T. 2001, *Inflectional Morphology. A Theory of Paradigm Structure*, Cambridge University Press, Cambridge, Royaume-Uni.

Stump, G. T. 2006, « Paradigm function morphology », dans *Encyclopedia of Language and Linguistics (2nd ed.)*, édité par K. Brown, Elsevier, Oxford, Royaume-Uni, p. 171–173.

Suchanek, F. M., G. Kasneci et G. Weikum. 2008, « Yago: A large ontology from Wikipedia and WordNet », *Journal of Web Semantics*, vol. 6, nº 3, p. 203–217.

Suignard, P. et S. Kerroua. 2013, « Utilisation de contextes pour la correction automatique ou semi-automatique de réclamations clients », dans *Actes de la 20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, France.

Swadesh, M. 1971, *The Origin and Diversification of Language (edited* post mortem *by Joel Sherzer)*, Aldine, Chicago, Illinois, États-Unis.

Tadić, M. 2007, « Building the croatian dependency treebank: the initial stages », *Suvremena lingvistika*, vol. 63, p. 85–92.

Tanaka-Ishii, K. 2005, « Entropy as an Indicator of Context Boundaries: An Experiment Using a Web Search Engine », dans *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, Jeju, République de Corée, p. 93–105.

Tanguy, L. et N. Hathout. 2002, « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web », dans *Actes de la 9ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2002)*, Nancy, France, p. 245–254.

Tavčar, A., D. Fišer et T. Erjavec. 2012, « sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic », dans *Proceedings of the 8th Language Technologies Conference, within the proceedings of the 15th International Multiconference Information Society (IS 2012)*, vol. C, Ljubljana, Slovénie, p. 197–202.

Telljohann, H., E. W. Hinrichs, S. Kübler, H. Zinsmeister et K. Beck. 2009, « Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z) », cahier de recherche, Seminar für Sprachwissenschaft, Universität Tübingen.

Tesnière, L. 1934, « Comment construire une syntaxe », *Bulletin de la Faculté des Lettres de Strasbourg*, vol. 12, nº 7, p. 219–229.

Tesnière, L. 1959, *Éléments de syntaxe structurale*, Klinksieck, Paris, France.

Thackston, W. M. 2006, « Kurmandji Kurdish: A reference grammar with selected readings », Http ://www.fas.harvard.edu/ iranian/Kurmanji/kurmanji_1_grammar.pdf.

Thomasset, F. et É. Villemonte de La Clergerie. 2005, « Comment obtenir plus des méta-grammaires », dans *Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, Dourdan, France.

Thornton, A. M. 2011, « Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in italian verb morphology », dans *Morphological Autonomy : Perspectives From Romance Inflectional Morphology*, édité par M. G. Martin Maiden, John Charles Smith et M.-O. Hinzelin, Oxford University Press.

Tiedemann, J. 2003, *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*, Thèse de doctorat, Uppsala universitet, Uppsala, Suède. Studia Linguistica Upsaliensia 1.

Tiedemann, J. et P. Nabende. 2009, « Translating transliterations », *International Journal of Computing and ICT Research, Special Issue*, p. 33–41.

Tolone, E. 2011, *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*, Thèse de doctorat, Université Paris-Est, France.

Tolone, E. et B. Sagot. 2009, « Using Lexicon-Grammar tables for French verbs in a large-coverage parser », dans *Proceedings of the 4th Language and Technology Conference (LTC 2009)*, édité par Z. Vetulani, Poznań, Pologne, p. 200–204.

Tolone, E. et B. Sagot. 2011, « Using Lexicon-Grammar tables for French verbs in a large-coverage parser », dans *Human Language Technology, Forth Language and Technology Conference (LTC 2009), Revised Selected Papers*, édité par Z. Vetulani, Lecture Notes in Artificial Intelligence (LNAI), Springer-Verlag, Poznań, Pologne.

Tolone, E., B. Sagot et É. Villemonte de La Clergerie. 2012, « Evaluating and improving syntactic lexica by plugging them within a parser », dans *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, La Valette, Malte.

Tolone, E., É. Villemonte de La Clergerie et B. Sagot. 2011, « Évaluation de lexiques syntaxiques par leur intégration dans l'analyseur syntaxique FRMG », dans *Proceedings of the 30th International Conference on Lexis and Grammar*, Nicosie, Chypre, p. 267–274.

Tosco, M. 2001, *The Dhaasanac language*, Rüdiger Klöppe, Cologne, Allemagne.

Tournier, J. 1988, *Précis de Lexicologie Anglaise*, Nathan, Paris, France.

Toutanova, K. et C. D. Manning. 2000, « Enriching the knowledge sources used in a maximum entropy part-of-speech tagger », dans *Proceedings of International Conference on New Methods in Language Processing*, Hong Kong, Chine, p. 63–70.

Toutanova, K. et R. C. Moore. 2002, « Pronunciation Modeling for Improved Spelling Correction », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, Philadelphie, Pennsylvanie, États-Unis, p. 144–151.

Tribout, D., A.-L. Ligozat et D. Bernhard. 2012, « Constitution automatique d'une ressource morphologique : VerbAgent », dans *Actes du 3ème Congrès Mondial de Linguistique Française (CMLF 2012)*, Lyon, France.

Troyanski, P. P. 1935, « Машина для подбора и печатания слов при переводе с одного языка на другой или на нескольких языках одновременно (Une machine pour la sélection et l'impression de mots pour la traduction d'une langue vers une autre ou vers plusieurs autres simultanément) », Brevet de l'Union des Républiques Socialistes Soviétiques n°40995.

T'sou, B. K., H.-L. Lin, G. Liu, T. Chan, J. Hu, C.-h. Chew et J. K. Tse. 1997, « A synchronous Chinese language corpus from different speech communities: Construction and applications », *Computational Linguistics and Chinese Language Processing*, vol. 2, n° 1, p. 91–104.

Tufiş, D. 2000, « BalkaNet – Design and Development of a Multilingual Balkan WordNet », *Romanian Journal of Information Science and Technology*, vol. 7, n° 1–2.

Tufiş, D., S. Koeva, T. Erjavec, M. Gavrilidou et C. Krstev. 2009, « Building language resources and translation models for machine translation focused on south Slavic and Balkan languages », dans *Scientific results of the SEE-ERA.NET Pilot Joint Call*, édité par J. Machačová et K. Rohsmann, p. 37–48.

Turcato, D., J. Toole, S. Tsiplakou, T. Heift et P. McFetridge. 2000, « An approach to lexical development for inflectional languages », dans *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athènes, Grèce.

Turian, J., L. Ratinov et Y. Bengio. 2010, « Word representations : A simple and general method for semi-supervised learning », dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Suède, p. 384–394.

Urieli, A. 2014, « Améliorer l'étiquetage de "que" par les descripteurs ciblés et les règles », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France, p. 56–66.

Urieli, A. et L. Tanguy. 2013, « L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane », dans *Actes de la 20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables d'Olonne, France.

Ushioda, A., D. Evans, T. Gibson et A. Waibel. 1993, « The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora », dans *Proceedings of the Workshop on the Acquisition of Lexical Knowledge from Text*, Columbus, Ohio, États-Unis, p. 95–106.

Vanrullen, T., P. Blache, C. Portes, S. Rauzy, J.-F. Maeyhieux, M.-L. Guénot, M.-L. Balfourier et J.-M. Bellengier. 2005, « Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales », dans *Actes de la 12ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, p. 41–48.

Vauquois, B., G. Veillon et J. Veyrunes. 1965, « Application des grammaires formelles aux modèles linguistiques en traduction automatique », *Kybernetika*, vol. 1, nº 3, p. 281–289.

Veiga, A., S. Candeias et F. Perdigão. 2013, « Generating a pronunciation dictionary for european portuguese using a joint-sequence model with embedded stress assignment », *Journal of the Brazilian Computer Society*, vol. 19, nº 2, p. 127–134.

Vernerová, A., V. Kettnerová et M. Lopatkova. 2014, « To Pay or to Get Paid: Enriching a Valency Lexicon with Diatheses », dans *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Islande.

Véronis, J. 1988, « Computerized correction of phonographic errors », *Computers and the Humanities*, vol. 22, nº 1, p. 43–56.

Veronis, J. 1998, « Multext-Lexicons, A set of Electronic Lexicons for European Languages », .

Versley, Y. et I. Rehbein. 2009, « Scalable discriminative parsing for German », dans *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*, Paris, France, p. 134–137.

Vijay-Shanker, K., D. Weir et A. K. Joshi. 1987, « Characterizing structural descriptions produced by various grammatical formalisms », dans *Proceedings of the 25th Meeting of the Association for Computational Linguistics (ACL'87)*, Stanford, Californie, États-Unis, p. 104–111.

Vilar, D., J.-T. Peter et H. Ney. 2007, « Can we translate letters? », dans *Proceedings of WMT 2007*, p. 33–39.

Villemonte de La Clergerie, É. 2005, « From metagrammars to factorized TAG/TIG parsers », dans *Proceedings of IWPT'05 (poster)*, Vancouver, Colombie Britannique, Canada, p. 190–191.

Villemonte de La Clergerie, É. 2013, « Improving a symbolic parser through partially supervised learning », dans *Proceedings of the 13th International Conference on Parsing Technologies (IWPT 2013)*, Naria, Japon.

Villemonte de La Clergerie, É. 2014, « Jouer avec des analyseurs syntaxiques », dans *Actes de la 21ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, Marseille, France.

Villemonte de La Clergerie, É., B. Sagot, L. Nicolas et M.-L. Guénot. 2009a, « FRMG : évolutions d'un analyseur syntaxique TAG du français », dans *Journée ATALA « Quels analyseurs syntaxiques pour le français ? » (organisée conjointement à la conférence IWPT 2009)*, édité par É. Villemonte de la Clergerie et P. Paroubek, Paris, France.

Villemonte de La Clergerie, É., B. Sagot et D. Seddah. 2017, « The ParisNLP entry at the CoNLL UD Shared Task 2017 : A Tale of a #parsingtragedy », dans *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Colombie Britannique, Canada, p. 243–252.

Villemonte de La Clergerie, É., B. Sagot, R. Stern, P. Denis, G. Recourcé et V. Mignot. 2009b, « Extracting and visualizing quotations from news wires », dans *Proceedings of the 4th Language and Technology Conference (LTC 2009)*, vol. 6562, édité par Z. Vetulani, Springer-Verlag, Poznań, Pologne, p. 522–532.

Vincze, V., J. Zsibrita et T. István Nagy. 2013, « Dependency parsing for identifying hungarian light verb constructions », dans *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, Nagoya, Japon, p. 207–215.

Vossen, P., éd.. 1999, *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer Academic Publisher, Dordrecht, Pays-Bas.

Vulić, I. et A. Korhonen. 2016, « Is "Universal Syntax" Universally Useful for Learning Distributed Word Representations ? », dans *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Allemagne, p. 518–524.

Walther, G. 2011a, « Latin passive morphology revisited », dans *Communication orale au colloque de la la Linguistic Association of Great-Britain (LAGB 2011)*, Manchester, Royaume-Uni.

Walther, G. 2011b, « Measuring morphological canonicity », *Linguistica*, vol. 51, p. 157–180. Internal and External Boundaries of Morphology.

Walther, G. 2013a, « Controlling arbitrariness: descriptive economy as an index of inflectional complexity », dans *Communication orale au 2nd American International Morphology Meeting (AIMM 2013)*, San Diego, Californie, États-Unis.

Walther, G. 2013b, *De la canonicité en morphologie — perspectives empiriques, formelles et computationnelles*, Thèse de doctorat, Université Denis–Diderot Paris 7, Paris, France.

Walther, G. 2016, « Paradigm realisation and the lexicon », dans *Morphological paradigms and functions*, édité par F. Kiefer, J. P. Blevins et H. Bartos, Brill, Leyde, Pays-Bas.

Walther, G., A. Antonov et G. Jacques. 2014a, « Defining direct/inverse systems: a canonical approach », dans *Communication orale au 16th International Morphology Meeting (IMM 16)*, Budapest, Hongrie.

Walther, G., G. Jacques et B. Sagot. 2013, « Uncovering the inner architecture of Khaling verbal morphology », Presentation at the 3rd Workshop on Sino-Tibetan Languages of Sichuan.

Walther, G., G. Jacques et B. Sagot. 2014b, « The opacity-compactness tradeoff: Morphomic features for an economical account of Khaling verbal inflection », dans *Communication orale au 16th International Morphology Meeting (IMM 16)*, Budapest, Hongrie.

Walther, G. et L. Nicolas. 2011, « Enriching morphological lexica through unsupervised derivational rule acquisition », dans *Proceedings of the International Workshop on Lexical Recources (WoLeR 2011)*, Ljubljana, Slovénie.

Walther, G. et B. Sagot. 2010, « Developing a Large-Scale Lexicon for a Less-Resourced Language: General Methodology and Preliminary Experiments on Sorani Kurdish », dans *Proceedings of the 7th SaLTMiL Workshop on Creation and use of basic lexical resources for less-resourced languages (LREC 2010 Workshop)*, La Valette, Malte.

Walther, G. et B. Sagot. 2011a, « Modélisation et implémentation de phénomènes flexionnels non-canoniques », *Traitement Automatique des Langues*, vol. 52, n° 2, p. 91–122.

Walther, G. et B. Sagot. 2011b, « Problèmes d'intégration morphologique d'emprunts d'origine anglaise en français », dans *Proceedings of the 30th International Conference on Lexis and Grammar*, Nicosie, Chypre.

Walther, G., B. Sagot et K. Fort. 2010, « Fast Development of Basic NLP Tools: Towards a Lexicon and a POS Tagger for Kurmanji Kurdish », dans *Actes du 29e Colloque sur le Lexique et la Grammaire*, Belgrade, Serbie.

Wang, H., J. Zhu, S. Tang et X. Fan. 2011, « A new unsupervised approach to word segmentation », *Computational Linguistics*, vol. 37, n° 3, p. 421–454.

Wehrli, E. 2007, « Fips, a "deep" linguistic multilingual parser », dans *Proceedings of the Workshop on Deep Linguistic Processing (DeepLP 2007)*, Prague, République Tchèque, p. 120–127.

Wong, S. H. S. 2004, « Fighting arbitrariness in wordnet-like lexical databases – a natural language motivated remedy », dans *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC 2004)*, Brno, République tchèque, p. 234–241.

Wurzel, W. U. 1984, *Flexionsmorphologie und Natürlichkeit*, Akademie Verlag, Berlin, République Démocratique Allemande.

Xanthos, A. 2008, *Apprentissage automatique de la morphologie — Le cas des structures racine-schème*, *Sciences pour la Communication*, vol. 48, Peter Lang, Berne, Suisse.

Xu, W., J. Tetreault, M. Chodorow, R. Grishman et L. Zhao. 2011, « Exploiting Syntactic and Distributional Information for Spelling Correction with Web-Scale N-gram Models », dans *Proceedings of EMNLP 2011*, p. 1291–1300.

Xue, N. 2003, « Chinese word segmentation as character tagging », *Computational Linguistics and Chinese Language Processing*, vol. 8, n° 1, p. 29–48.

Yarowsky, D., G. Ngai et R. Wicentowski. 2001, « Inducing multilingual text analysis tools via robust projection across aligned corpora », dans *Proceedings of the Human Language Technology Conference (HLT 2001)*, San Diego, Californie, États-Unis.

Yngve, V. H. 1955, « Syntax and the problem of multiple meaning », dans *Machine Translation of Languages : Fourteen Essays*, édité par W. N. Locke et A. D. Booth, MIT Press, Cambridge, Massachusetts, États-Unis, p. 208–226.

Yokoi, T. 1995, « The EDR electronic dictionary », *Communications of the ACM*, vol. 38, n° 11, p. 42–44.

Younger, D. H. 1967, « Recognition and parsing of context-free languages in time n^3 », *Information and Control*, vol. 10, n° 2, p. 189–208.

Yu, S., H.-m. Duan, X.-f. Zhu et B. Sun. 2002a, « The specification of basic processing of contemporary Chinese corpus », *Journal of Chinese Information Processing*, vol. 16, n° 5, p. 49–64.

Yu, S., H.-m. Duan, X.-f. Zhu et B. Sun. 2002b, « The specification of basic processing of contemporary Chinese corpus (continued) », *Journal of Chinese Information Processing*, vol. 16, n° 6, p. 58–64.

Yvon, F. 2011, « `spellChecker` : un système de correction automatique fondé sur des automates probabilistes », Livrable du Projet TRACE (ANR-09-CORD-023). Accessible à l'adresse `http://anrtrace.limsi.fr/dev/Anr_trace_-_lot3.pdf`.

Zabokrtský, Z. et M. Lopatková. 2007, « Valency Information in VALLEX 2.0: Logical Structure of the Lexicon », *The Prague Bulleltin of Mathematical Linguistics*, vol. 87, p. 41–60.

Zanchetta, E. et M. Baroni. 2005, « Morph-it! a free corpus-based morphological resource for the Italian language », dans *Proceedings of the Corpus linguistics Conference*, Birmingham, Royaume-Uni, p. 1–12.

Zeman, D. 2002, « Can subcategorization help a statistical dependency parser? », dans *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, Taipei, Taïwan, p. 1–7.

Zeman, D., M. Popel, M. Straka, J. Hajic, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinkova, J. Hajic jr., J. Hlavacova, V. Kettnerová, Z. Uresova, J. Kanerva, S. Ojala, A. Missilä, C. D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, c. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. Elkahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. F. Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonca, T. Lando, R. Nitisaroj et J. Li. 2017, « CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies », dans *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Colombie Britannique, Canada, p. 1–19.

Zhao, H. et Q. Liu. 2010, « The CIPS-SIGHAN CLP 2010 Chinese word segmentation bakeoff », dans *Proceedings of the First CPS-SIGHAN Joint Conference on Chinese Language Processing*, Pékin, Chine, p. 199–209.

Zhikov, V., H. Takamura et M. Okumura. 2010, « An efficient algorithm for unsupervised word segmentation with branching entropy and MDL », dans *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, Cambridge, Massachusetts, États-Unis, p. 832–842.

Zielinski, A. et C. Simon. 2009, « Morphisto – an open source morphological analyzer for German », dans *Post-proceedings of the 7th conference on Finite-State Methods and Natural Language Processing (FSMNLP 2008)*, IOS Press, Amsterdam, Pays-Bas, p. 224–231.

Ziv, J. et A. Lempel. 1977, « A universal algorithm for sequential data compression », *IEEE Transactions on Information Theory*, vol. 23, nº 3, p. 337–343.

Zúñiga, F. 2006, *Deixis and Alignment - Inverse systems in indigenous languages of the Americas*, Benjamins, Amsterdam, Pays-Bas.

Zwicky, A. M. 1985, « How to Describe Inflection », dans *Proceedings of the 11th Annual Meeting of the Berkeley Linguistics Society*, Berkeley, Californie, États-Unis, p. 372–386.

Zwicky, A. M. et G. K. Pullum. 1988, « The syntax-phonology interface », dans *Linguistic Theory : Foundations*, édité par F. J. Newmeyer, chap. 10, Linguistics : The Cambridge Survey, Cambridge University Press, Cambridge, Royaume-Uni, p. 255–280.